# Modelo de medición de riesgo crediticio aplicado a cooperativa multiactiva

Por

**Melissa García Herrera**

**Lina Marcela Aguilar Narváez**

Trabajo presentado como requisito parcial para optar al título de
Magíster en Administración Financiera

Asesor
Brayan Ricardo Rojas Ormaza, FRM, Msc

UNIVERSIDAD EAFIT

Santiago de Cali, marzo 2022

## Resumen

La gestión del riesgo de crédito es un enfoque estructurado que las organizaciones solidarias han adoptado para identificar, medir, monitorear y controlar de manera efectiva el riesgo, de modo que puedan tomar decisiones oportunas para prevenir y reducir el incumplimiento en el pago de las obligaciones contraídas por los asociados. El presente trabajo tiene como objetivo diseñar un modelo de medición de riesgo crediticio que, a su vez, cumpla con los lineamientos impartidos por el ente regulatorio y que permita estandarizar el proceso de otorgamiento de créditos para una cooperativa multiactiva, con el fin de facilitar la toma de decisiones, especialmente de los comités de crédito y evaluación de cartera. Para ello, se utilizaron los modelos de Regresión logística, Árboles de decisión y Random Forest, con el fin de determinar desde la eficiencia el modelo más adecuado para esta organización.



**Palabras claves**: Riesgo crediticio, Asociados, Modelos, Gestión del riesgo.



## Abstract

Credit risk management is a structured approach that solidarity organizations have adopted to identify, measure, monitor and effectively control risk, so that they can take timely decisions to prevent and reduce the default in the payment of obligations by partners. The objective of this paper is to design a model for measuring credit risk to comply with the guidelines given by the regulatory entity and allows standardizing the process of granting credit to a multi-active cooperative in order to help in decision-making, especially by the credit and portfolio evaluation committees. For this, Logistic Regression, Decision Tree and Random Forest models were used, in order to determine from efficiency the most appropriate model for this organization.



**Key words:** Credit risk, Associates, Models, Risk management.

## Contenido


1. Introducción ..... 5
2. Situación de estudio ..... 7
3. Objetivos ..... 9
   3.1 Objetivo general ..... 9
   3.2 Objetivos específicos ..... 9
4. Marco teórico ..... 10
   4.1 Modelo Logit ..... 12
   4.2 Árboles de decisión ..... 14
      4.2.1 Chaid (Chi-square automatic interaction detector): ..... 16
      4.2.2 Árbol C5.0 ..... 16
   4.3 Random Forest ..... 17
   4.4 Modelo regulatorio ..... 19
5. Pruebas de bondad de ajuste ..... 25
   5.1 Curva ROC ..... 25
   5.2 Test de Hosmer-Lemeshow ..... 26
   5.3 Matriz de confusión ..... 26
   5.4 Prueba de Kolmogorov-Smirnov o K-S ..... 27
6. Metodología ..... 28
7. Resultados ..... 29
   7.1 Desarrollo del modelo ..... 29
      7.1.1 Variables ..... 30
   7.2 Análisis de datos en software ..... 31
      78.2.1 Modelo de regresión logística binaria ..... 31
      7.2.2 Modelo árboles de decisión ..... 36
      7.2.3 Random Forest ..... 40
   7.3 Comparación entre los modelos ..... 45
8. Conclusiones y recomendaciones ..... 48

Referencias ..... 50

Anexos ..... 53

## Índice de tablas

Tabla 1. Categorías de riesgos ..... 11
Tabla 2. Ventajas y desventajas de los modelos (Logit, Árboles de Decisión y Random Forest) ..... 21
Tabla 3. Estudios relacionados con el tema tratado ..... 23
Tabla 4. Matriz de confusión ..... 26
Tabla 5. Frecuencia por clase muestral ..... 30
Tabla 6. Variables en la ecuación-Modelo Logit ..... 32
Tabla 7. Elección del punto de corte ..... 34
Tabla 8. Matriz de confusión-Modelo Logit ..... 34
Tabla 9. Perfiles de riesgo ..... 36
Tabla 10. Elección punto de corte ..... 38
Tabla 11. Matriz de confusión- árboles de decisión ..... 39
Tabla 12. Elección de puntos de corte – modelos de Random Forest ..... 42
Tabla 13. Matriz de confusión- modelos de Random Forest ..... 43
Tabla 14. Comparación entre modelos ..... 45
Tabla 15. Rangos de calificación por modelo ..... 46
Tabla 16. Cálculo del deterioro para la modalidad de consumo ..... 47

## Índice de figuras

Figura 1. Etapas del crédito ..... 7
Figura 2. Gráfico de la función sigmoidea ..... 14
Figura 3. Diagrama árbol de decisión ..... 15
Figura 4. Diagrama de Random Forest ..... 17
Figura 5. Curva de ROC ..... 25
Figura 6. Curva de ROC-Modelo Logit ..... 35
Figura 7. Diagrama modelo árbol de decisión ..... 37
Figura 8. Curva ROC- árboles de decisión ..... 40
Figura 9. Proceso one-hot encoding ..... 41
Figura 10. Curva ROC-Modelos Random Forest ..... 44
Figura 11. Curva ROC-Comparación Modelos ..... 45

# 1. Introducción

El análisis de riesgo de crédito es un tema que ha tomado relevancia en las organizaciones, toda vez que, a través de técnicas analíticas, evaluaciones cualitativas o el uso de ambas, se pueden determinar las variables o factores (edad, género, capacidad de pago, entre otras) que explican el por qué un individuo posee una mayor probabilidad de incumplir con sus acreencias que otro. A su vez, hace parte de los marcos técnicos de información financiera, así como las buenas prácticas para la adecuada administración del riesgo crediticio que en su conjunto conforman el Sistema de Administración de Riesgo Crediticio (SARC). De acuerdo con el capítulo II de Circular Básica Contable y Financiera de la Superintendencia de la Economía Solidaria (2020): "El SARC debe contener políticas y procedimientos claros y precisos que definan los criterios y la forma mediante la cual la organización solidaria vigilada identifica, evalúa, asume, califica, controla y cubre su riesgo crediticio" (p. 1).

La cooperativa multiactiva, objeto de este estudio, cuenta con más de 30 años de historia ofreciendo servicios y beneficios a egresados, directivos, profesores y demás trabajadores de la Universidad Autónoma De Occidente, contribuyendo así al mejoramiento de la calidad de vida de estos y sus familias. No obstante, dentro del Plan de Desarrollo tiene como meta ofrecer sus servicios en otras empresas para incluir asociados que no pertenecen a la universidad. Lo anterior supone una mayor exposición al riesgo de crédito que debe ser mitigada a través de un sistema de gestión de riesgo, que permita la toma de decisiones adecuadas y oportunas. Por esta razón, desde el Comité de Crédito y Evaluación de Cartera se ha detectado la necesidad de implementar un modelo para responder a este reto organizacional.

Acorde con la situación que enfrenta la organización, el propósito de este trabajo es desarrollar dicho modelo para estandarizar el proceso de otorgamiento de créditos y el seguimiento de los mismos, facilitando la toma de decisiones al comité evaluador.

Para el caso particular de esta cooperativa, la utilización de los modelos Logit, árboles de decisión y Random Forest clasificaron principalmente a los clientes malos con las siguientes características: ser asociados con pocos aportes, ser ex empleados de la universidad, tener niveles de ingresos bajos y tener créditos con montos altos. Cabe mencionar que la base de datos presentó diferentes retos como son: la consolidación de la información consignada en los formularios de inscripción versus la depositada en el sistema, lo cual dificultó la definición de las variables a considerar dentro de los modelos y además la cantidad de asociados incumplidos era muy poca, por eso fue necesario ponderar para balancear la base de datos y posteriormente aplicar nuevamente la regresión logística binaria y otros modelos como: árboles de decisión y Random Forest para estimar la probabilidad de incumplimiento.

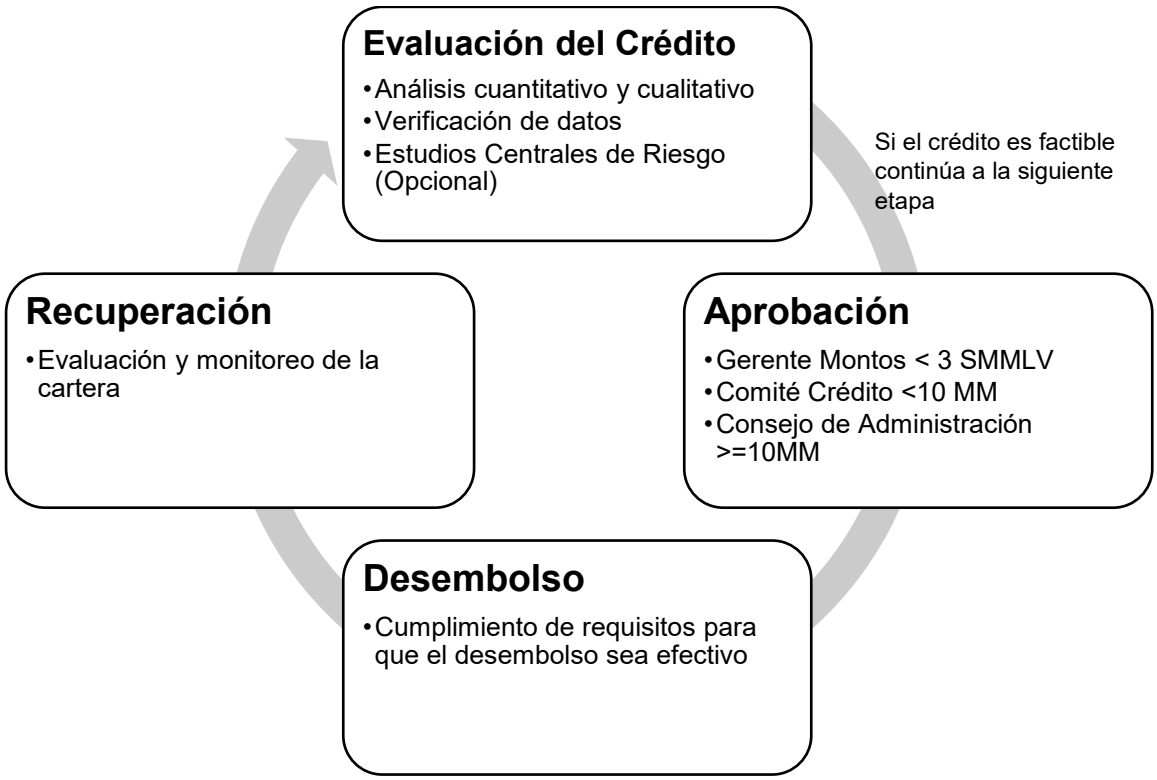
Los resultados obtenidos a partir de los modelos estadísticos se contrastaron con los datos que actualmente maneja la cooperativa de sus asociados, lo que permitirá confirmar o debatir qué tan acertado es el manejo del riesgo de crédito de la entidad.

Para el desarrollo de los objetivos propuestos, este estudio se organiza en las siguientes secciones: la primera es el actual componente; la segunda se divide en dos partes: a) la revisión de literatura existente para modelos de riesgo de crédito: regresión logística binaria, árboles de decisión y Random Forest, b) las principales pruebas de bondad de ajuste, seguido a esto, se hace una especificación metodológica que desagrega el paso a paso para la estimación del perfil de riesgos de la cooperativa, luego, se exponen los resultados del modelo, y finalmente, se hace una presentación de las conclusiones y recomendaciones para la organización.

## 2. Situación de estudio

Actualmente la cooperativa, no posee una herramienta que estime la probabilidad de incumplimiento del asociado dentro del proceso de aprobación y desembolso de los créditos. Este proceso se realiza de la siguiente forma:

**Figura 1.** Etapas del crédito




Fuente: Elaboración propia, 2022.

Cabe mencionar que el estudio en centrales de riesgo se lleva a cabo porque es una información que genera valor agregado en el proceso de otorgamiento y contribuye a proteger el patrimonio de los asociados, además de mantener una sana condición de la cartera.

Aunque el riesgo se mitiga en gran medida con la vinculación de asociados que pertenecen a la universidad, presentan capacidad de endeudamiento, años de

antigüedad requeridos, garantías constituidas y el recaudo está garantizado a través del descuento por nómina, al ampliar la oferta de créditos a asociados no vinculados a la universidad; se debe hacer uso del recaudo por caja, siendo necesario realizar una medición del riesgo de crédito, no solo al momento del otorgamiento sino en etapas posteriores, dada la evolución de las condiciones económicas de los mismos, lo cual trae consigo un potencial deterioro en la cartera.

Tal como lo menciona la Circular Básica Contable y Financiera de la Superintendencia de la Economía Solidaria (2020), la implementación del sistema de administración del riesgo de crediticio (SARC), contempla varias etapas, entre las que se tiene la etapa de identificación, donde mínimamente:

> […] se debe realizar la identificación de los riesgos de crédito, cuando se presenten las siguientes situaciones: i) previo al lanzamiento, uso o modificación de cualquier producto; ii) antes de la inclusión de nuevas prácticas comerciales, incluyendo nuevos canales de prestación de servicios, y el uso de nuevas tecnologías o tecnologías en desarrollo para productos nuevos o existentes; iii) previo a la incursión en un nuevo mercado o nicho de mercado. (p. 5)

El modelo propuesto permitirá estimar la probabilidad de incumplimiento del solicitante del crédito, involucrando variables cualitativas y cuantitativas, que aunadas a los demás criterios técnicos empleados actualmente por la entidad, faciliten a los comités de crédito la toma de decisiones y permita cumplir con las disposiciones de la Superintendencia de la Economía Solidaria en lo concerniente a la administración del riesgo de crédito.

Es importante precisar que el modelo de referencia para el cálculo de la pérdida esperada, propuesto por la Superintendencia, debe ser aplicado por las cooperativas de ahorro y crédito, multiactivas e integrales con sección de ahorro y crédito, los fondos de Empleados de categoría plena y las demás organizaciones

solidarias que adelanten actividad crediticia del primer y segundo nivel de supervisión. Al no cumplir con las condiciones anteriormente mencionadas, la cooperativa continuará constituyendo el deterioro por altura de mora con factor de riesgo, según las disposiciones contenidas en el Anexo 1 del Capítulo II y trabajando en distintas opciones de modelación, con el objetivo de prever posibles cambios en normatividad.

# 3. Objetivos

## 3.1 Objetivo general

Construir un modelo de medición de riesgo crediticio, que facilite la toma de decisiones al comité evaluador en el proceso de otorgamiento de créditos y el seguimiento de los mismos.

## 3.2 Objetivos específicos

- Identificar las variables que inciden en el proceso actual de otorgamiento de los créditos.
- Analizar los datos para hacer uso de la metodología adecuada.
- Construir un modelo que permita evaluar el riesgo crediticio de los asociados y esté acorde con los lineamientos del ente regulador.

## 4. Marco teórico

Las cooperativas juegan un papel importante en la economía porque son impulsoras de la generación de empleo, equidad social e igualdad. Toda vez que en estas organizaciones prevalecen los intereses sociales sobre los económicos, contribuyendo así a la construcción de confianza y tejido social en las comunidades.

Este tipo de empresa asociativa sin ánimo de lucro, donde los trabajadores o los usuarios son simultáneamente los aportantes y los gestores, está vigilada por la Superintendencia de la Economía Solidaria, ente encargado de la vigilancia y control de la actividad financiera del cooperativismo y los servicios de ahorro y crédito de los fondos de empleados y asociaciones mutuales.

Para el cumplimiento de dichas funciones, la Supersolidaria dispone de un marco regulatorio que comprende leyes, decretos, acuerdos, circulares y resoluciones que reglamentan las actuaciones de la entidad en el desarrollo de su objeto misional. Recientemente actualizó la Circular Básica Contable y Financiera por medio de la Circular Externa 22 de 2020, la cual sustituye y deroga la Circular Básica Contable y Financiera 04 de 2008 y todas las Circulares Externas y Cartas Circulares que resulten contrarias, con el objetivo de alinear los marcos técnicos normativos de información financiera, normas sobre regulación prudencial para las cooperativas de ahorro y crédito, y las buenas prácticas para la adecuada administración o gestión de riesgos, con las nuevas disposiciones que el Gobierno Nacional ha expedido para el sector de la Economía Solidaria.

En el desarrollo de este ejercicio compete revisar el capítulo II de la Circular Básica Contable y Financiera, el cual establece los parámetros para la medición del riesgo de crédito de las entidades vigiladas en lo que denomina el Sistema de Administración de Riesgo de Crédito (SARC). Las organizaciones solidarias están expuestas a este tipo de riesgo, tanto al momento de otorgar créditos como a lo

largo de la vida de los mismos, por lo que dicho sistema deberá permitirles adoptar decisiones oportunas para la adecuada mitigación del riesgo crediticio y evitar incurrir en pérdidas que amenacen la solidez patrimonial, definiendo el riesgo crediticio como: “la probabilidad de que una organización solidaria incurra en pérdidas y disminuya el valor de sus activos, como consecuencia del incumplimiento del pago de las obligaciones contractuales por parte de sus deudores o contraparte” (Supersolidaria, 2020, p. 118).

Para llevar a cabo la medición de riesgo de crédito, se deberán tener en cuenta dos variables que se definen en el capítulo II de la Circular Básica Contable y Financiera:

**Probabilidad de incumplimiento del deudor (PI):** Probabilidad de que un asociado incumpla parcial o totalmente una obligación en un horizonte de tiempo, explicado por un conjunto de variables sociodemográficas, historial crediticio e indicadores financieros. Esta variable permitirá evaluar el perfil del asociado y clasificarlo en las siguientes categorías:

**Tabla 1.** Categorías de riesgos

| Categoría | Tipo de Riesgo |
|---|---|
| A | Riesgo Normal |
| B | Riesgo Aceptable, Superior al Normal |
| C | Riesgo Apreciable |
| D | Riesgo Significativo |
| E | Riesgo de Incobrabilidad |

Fuente: Anexo 1, Consideraciones para la calificación y deterioro de la cartera de crédito, Capitulo II Sistema de Administración del Riesgo de Crédito, Circular Básica Contable y Financiera, 2020.

**Severidad o Pérdida dado el incumplimiento (PDI):** Corresponde al porcentaje (%) promedio que no se logra recuperar de hacer efectiva la garantía ante un incumplimiento. La PDI varía por tipo de garantía.

Finalmente, la pérdida esperada (deterioro) para una determinada modalidad de crédito, se calcula aplicando la siguiente fórmula:

**Ecuación 1.** PÉRDIDA ESPERADA = [Probabilidad de incumplimiento] x [Valor expuesto del activo] x [Pérdida dado el incumplimiento]. [1]

Para la estimación de la probabilidad de incumplimiento y cuantificar las pérdidas esperadas es importante revisar modelos de *credit scoring*, los cuales son algoritmos que evalúan automáticamente el riesgo crediticio del solicitante del préstamo. Estos modelos se caracterizan por utilizar la información de las solicitudes de crédito, fuentes de información interna y/o externa, para asignar al evaluado un puntaje o *score* que se convierte en una calificación de riesgo que permite evaluar qué tan viable es otorgar un crédito y el riesgo asumido en dicha operación.

Alguna de las metodologías más utilizadas son las siguientes:

## 4.1 Modelo Logit

Este modelo tiene sus orígenes en la familia de los modelos de respuesta cualitativa, en donde la variable dependiente es de tipo binario, es decir, solo toma dos valores: 1 cuando el objeto de estudio pertenece a una condición específica (éxito) o 0 si es lo contrario (fracaso), facilitando el cálculo de la probabilidad que tiene un cliente de pertenecer a uno de los grupos establecidos a priori (no pagador o pagador). Por lo que una de sus ventajas es que al mantener la variable explicada en un rango entre cero y uno, puede ser empleado en la medición de la probabilidad de incumplimiento. La introducción de la logística como una opción válida se atribuye

[1] Modelos para la estimación o cuantificación de pérdidas esperadas, Capítulo II Circular Básica Contable y Financiera, Supersolidaria, 2020.

al físico y estadístico Joseph Berkson, quien en 1944 propuso su uso, acuñando el término 'logit' (Cramer, 2003).

El modelo Logit ha evolucionado de forma independiente en varias disciplinas, que van desde las ciencias de la salud hasta el marketing. Cabe mencionar que las variables independientes que utiliza la regresión logística pueden ser continuas o categóricas, pero a diferencia de la regresión lineal, la variable dependiente de la regresión logística toma un número limitado de categorías en lugar de un resultado continuo, es decir, son de tipo binario como se mencionó anteriormente.

La ecuación usada por dicho modelo es la función sigmoide:

**Ecuación 2.** $P_i = \frac{1}{1+e^{-(Z_i)}} = \frac{e^z}{1+e^z}$

Donde:

$P_i$: es la probabilidad de incumplimiento.

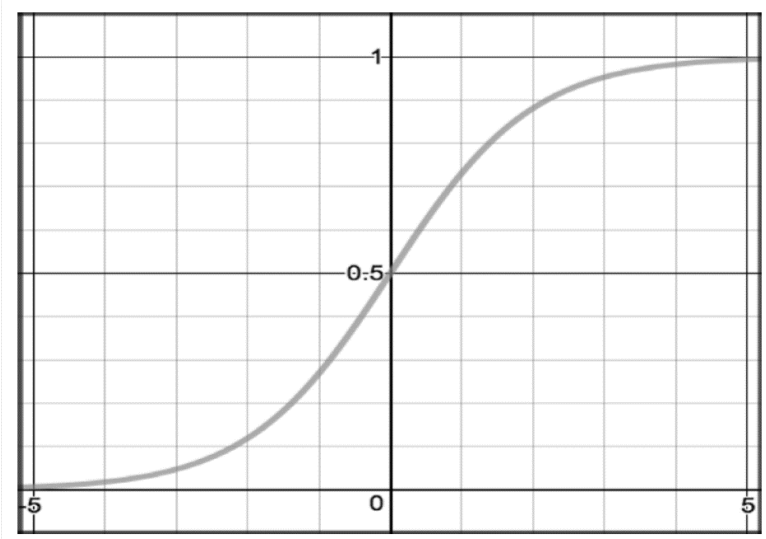
$Z_i$: es el scoring logístico y equivale a $z_i = \beta_1 + \beta_2 X_i$

e: es el número de Euler.

$Z_i$ es un número real proveniente de una regresión lineal; cuando z tiende a $\infty$ entonces el valor predicho de Pi se convertirá en 1 y si tiende a $-\infty$ el valor predicho de Pi se convertirá en cero. La metodología de este modelo parte de generar un número continuo que se interpreta como la ocurrencia o no de un evento, en términos de probabilidad.

Gráficamente tiene una forma de “S”, como se observa en la siguiente figura:

**Figura 2.** Gráfico de la función sigmoidea



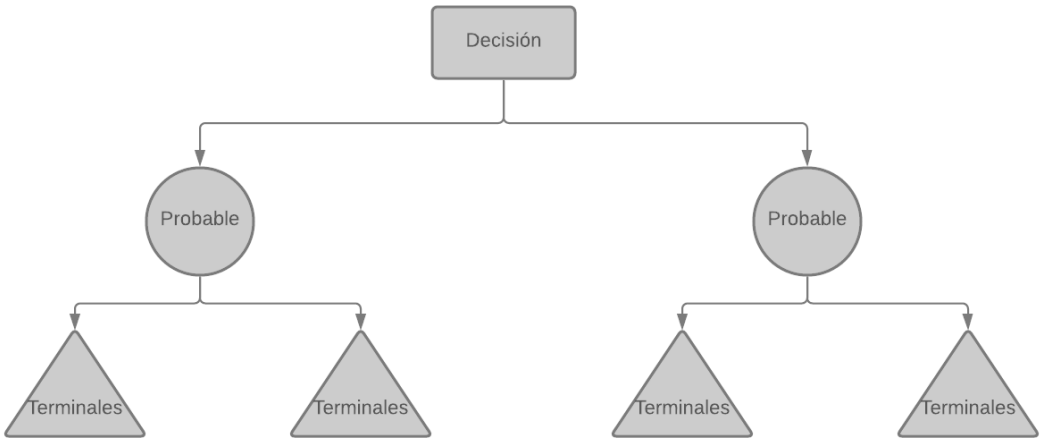

Fuente: ICHI PRO. ¿Qué es la función sigmoidea? ¿Cómo se implementa en regresión logística? Recuperado de https://ichi.pro/es/que-es-la-funcion-sigmoidea-como-se-implementa-en-regresion-logistica-77981969140323

## 4.2 Árboles de decisión

Metodología clave en el análisis de *credit score*, está clasificada dentro de las técnicas no paramétricas y su uso es muy común en diversos campos, tales como: economía, biología, sociología, ciencias de la computación, por mencionar algunos.

Un árbol de decisión es un método gráfico y analítico que se utiliza para representar eventos que pueden surgir de una decisión tomada en un momento determinado, dado un conjunto de decisiones posibles, que ayudan a tomar la decisión más certera desde el punto de vista probabilístico. Su componente gráfico permite establecer las relaciones entre variables y entender la estructura del modelo.

**Figura 3.** Diagrama árbol de decisión




Fuente: Elaboración propia, 2022.

Como se puede observar en la figura 3, un árbol de decisión contiene los siguientes elementos:

- Nodo de decisión: Indica la decisión a tomar y se representa con un cuadro.
- Nodo de probabilidad: Muestra la probabilidad de obtener cierto resultado y se representa con un círculo.
- Nodo terminal: Muestra el resultado definitivo de una ruta de decisión y se representa con un triángulo.

La importancia de emplear esta metodología en el análisis de *credit score* radica en que un mismo modelo permite diferentes usos, como el seguimiento de clientes considerados como buenos (probabilidades bajas de incumplimiento) y discriminar de acuerdo al nivel de riesgo, los clientes con mayor probabilidad de llegar a incumplimiento.

Los árboles tienen la ventaja de que no requieren la formulación de supuestos estadísticos sobre distribuciones estadísticas o formas funcionales. A su vez, presentan la relación entre las variables, los grupos y el riesgo de manera visual,

con lo cual, si el conjunto de variables en el análisis es reducido, facilita entender cómo funciona el *scoring* (Gutiérrez Girault, 2007, p. 5).

Para trabajar con árboles de decisión existen varias metodologías, entre las que se destacan:

### 4.2.1 Chaid (Chi-square automatic interaction detector)

Se basa en el algoritmo AID (Automatic Interaction Detection) o Detección Automática de Interacciones, llamado así porque inicialmente no se tenía como objetivo la clasificación, sino que estaban centrados en las interacciones entre las variables.

Consiste en un rápido y eficaz algoritmo de árbol estadístico y multidireccional, que explora datos creando segmentos y perfiles con respecto al resultado deseado. Permite la detección automática de interacciones mediante Chi-cuadrado en lugar del análisis de la varianza. La construcción del árbol se realiza eligiendo la variable independiente (predictora) que presenta la interacción más fuerte con la variable dependiente. Las categorías de cada predictor se funden si no son significativamente distintas respecto a la variable dependiente.

### 4.2.2 Árbol C5.0

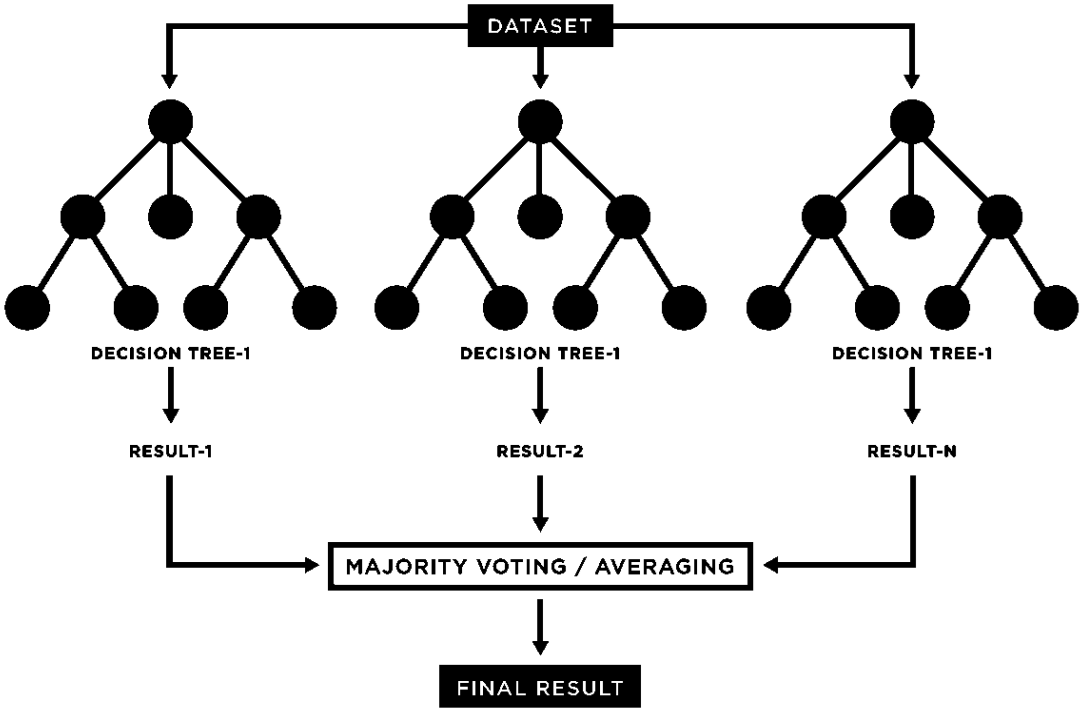
El algoritmo C5 y, sobre todo, su versión no comercial, C4.5, es uno de los más utilizados en el ámbito de los árboles de clasificación; fue desarrollado por Quinlan en 1993, como una extensión (mejora) del algoritmo ID3 que elaboró en 1986. Una de sus características es la división de la muestra en función del campo que ofrece la máxima ganancia de información. A su vez, las distintas submuestras definidas por la primera división se vuelven a dividir, por lo general basándose en otro campo, y el proceso se repite nuevamente hasta que resulta imposible dividir las

submuestras. Por último, se vuelven a examinar las divisiones del nivel inferior, y se eliminan o suprimen las que no contribuyen significativamente con el valor del modelo.

## 4.3 Random Forest

Los Random Forest o árboles aleatorios, inicialmente fueron propuesto por Tin Kam Ho bajo el nombre de Random decision forests en 1995, y después fueron desarrollados y llamados Random Forest por Leo Breiman (2001). Esta técnica de maching learning crea varios árboles de decisión sin correlación, que al combinar sus resultados, se obtiene un resultado final como lo muestra la siguiente figura.

**Figura 4.** Diagrama de Random Forest




Fuente: TIBCO Software. *What is a Random Forest?* TIBCO Software. Recuperado de https://www.tibco.com/reference-center/what-is-a-random-forest

La poca correlación entre los árboles generados por el Random Forest es la clave de esta metodología, ya que puede producir predicciones en conjunto, que son más precisas que cualquiera de las predicciones particulares, salvaguardando a los árboles entre sí de sus errores individuales.

Aunque algunos árboles del modelo pueden estar incorrectos, muchos otros están en lo cierto, por lo que, como grupo, en general estos árboles pueden estar moviéndose en la dirección correcta. Es por esto que el concepto esencial de esta técnica de aprendizaje automático supervisado es “Cien es mejor que uno”, dado que supera a cualquiera de los modelos constituyentes individuales, como ya se ha mencionado.

Cada uno de los árboles de decisión individuales se entrenan con muestras aleatorias extraídas de los datos originales mediante bootstrapping[2], es decir, cada árbol se entrena con datos ligeramente diferentes, y en cada árbol individual, las observaciones se distribuyen por nodos (igual a los árboles de decisión), generando la estructura del árbol hasta que se alcanza un nodo terminal, para después obtener la predicción que se obtiene del promedio o votación de las predicciones de todos los árboles individuales que componen el Random Forest.

Gracias a los buenos resultados que este modelo ha obtenido, se ha convertido en uno de los algoritmos que se ha utilizado en el sector financiero para segmentación de clientes, análisis de comportamiento de mercados, detección de fraude, gestión del riesgo, como la estimación de la probabilidad que un cliente no cumpla el pago de sus obligaciones financieras, entre otros.

[2] Este método estima la distribución de muestreo al tomar múltiples muestras con reemplazo de una sola muestra aleatoria. Estas nuevas muestras se denominan muestras repetidas. Cada muestra tiene el mismo tamaño que la muestra original. La muestra original representa la población de la cual se extrajo. Por lo tanto, las muestras repetidas de esta muestra original representan lo que obtendríamos si tomáramos muchas muestras de la población (¿Qué es un procedimiento bootstrap?, s.f.).

La transición de las entidades financieras hacia la digitalización ha impulsado el desarrollo de nuevos modelos, que utilizan herramientas como Python, software de código abierto, y tiene un lenguaje de fácil entendimiento que se ha convertido en el puente del entorno informático y empresarial, toda vez que conecta a ambos, posibilitando la facilidad de creación de códigos entendibles y de rápido aprendizaje, que son imprescindibles en modelos de *machine learning* como el Random Forest. Es preciso mencionar que Python emplea librerías[3] para el desarrollo de los modelos, entre las más conocidas se tiene Pandas[4], NumPy[5], Scikit-learn[6], entre otras.

## 4.4 Modelo regulatorio

El modelo regulatorio que deberán aplicar las organizaciones solidarias para el cálculo de la probabilidad de incumplimiento, de acuerdo a los lineamientos impartidos por la Supersolidaria, se basa en un modelo Logit, que calcula un puntaje de acuerdo a las características particulares de cada asociado y la modalidad de cartera a la cual pertenece (cartera consumo con libranza, cartera consumo sin libranza y cartera comercial-persona natural).

Con base en el puntaje obtenido, se establecen las calificaciones de riesgo, las cuales determinarán la PI correspondiente. Para la organización objeto de estudio, es importante revisar el modelo cartera consumo con libranza, por lo que el valor de Z (scoring logístico) para esta modalidad de cartera está dado por:

[3] Biblioteca que tiene un conjunto de implementos funcionales específicos para determinados usos.
[4] Librería de Python especializada en el manejo y análisis de estructuras de datos, como: cargar, modelar, analizar, manipular y prepararlos (Alberca, La librería Pandas, s.f.).
[5] Librería de Python especializada en el cálculo numérico y el análisis de datos, especialmente para un gran volumen de datos (Alberca, La librería Numpy, s.f.).
[6] Es una librería de código abierto que unifica bajo un único marco los principales algoritmos y funciones, facilitando en gran medida todas las etapas de preprocesado, entrenamiento, optimización y validación de modelos predictivos (Amat Rodrigo, 2020, Machine Learning con Python y Scikitlearn).

**Ecuación 3.** Modelo regulatorio cartera de consumo con libranza

$$Z = -2.2504 - 0.8444 * EA - 1.0573 * AP + 1.0715 * TC - 0.0139 * FE + 0.4187 * ESIN + 0.5313 * FAMOR - 0.5536 * VALCUOTA - 0.3662 * VALPRES + 0.0586 * OCOOP - 0.5981 * FONAHO - 1.3854 * COOCDAT - 0.5893 * FONDPLAZO + 0.7833 * ANTIPRE1 + 0.8526 * MORA15 + 1.4445 * MORA1230 + 1.3892 * MORA1260 + 0.2823 * MORA2430 + 0.7515 * MORA2460 - 0.6632 * SINMORA + 1.2362 * MORTRIM$$

Fuente: Anexo 2, modelos de referencia para la estimación de las pérdidas esperadas, Capitulo II Sistema de Administración del Riesgo de Crédito, Circular Básica Contable y Financiera, 2020.

Donde las variables toman el valor cero o uno, dependiendo si cumplen o no con la condición y, estas incluyen comportamiento de pago del asociado medido en distintas alturas de mora, así como otras características del asociado, tales como: la cantidad de aportes en la organización, tiempo de vinculación, si posee otros productos activos en la entidad, entre otras.

Luego de revisar los diferentes modelos, en el siguiente cuadro se sintetizan las principales ventajas y desventajas de los mismos:

**Tabla 2.** Ventajas y desventajas de los modelos (Logit, Árboles de Decisión y Random Forest)

| Modelo | Ventajas | Desventajas |
|---|---|---|
| **Logit** | *Este modelo lineal generalizado con distribución binomial, permite que los coeficientes se interpreten como una razón de probabilidad, facilitando la interpretación de los mismos.<br>* Dado que la variable dependiente del modelo es dicotómica no es necesario que se cumpla el supuesto de normalidad.<br>* En estos modelos es posible clasificar a un individuo en un grupo u otro, ya que con la función de enlace se obtiene un número binario, que permite hacer dicha clasificación. | * En este tipo de modelo el tamaño de muestra debe ser grande, ya que tiene como método de estimación la máxima verosimilitud.<br>* Este modelo logístico funciona bien para la predicción de resultados categóricos y multinomiales, sin embargo, este no puede predecir datos continuos.<br>* Incluir muchas variables independientes al modelo genera overfitting, lo cual causa que el modelo tenga poco valor predictivo, además de tener problemas como la multicolinealidad, ya que pueden existir variables relacionadas. |
| **Árboles de Decisión** | * Los árboles de decisión son simples de comprender e interpretar.<br>* No requieren una preparación de datos muy compleja.<br>* Se pueden aplicar tanto en variables cuantitativas como variables cualitativas, el tipo de dato no es una restricción.<br>* Su representación gráfica es fácil de entender.<br>* La respuesta del algoritmo puede entenderse fácilmente mediante el razonamiento lógico booleano aplicado en él. | * Construir un árbol de decisión con pocos datos (árbol pequeño) dará como resultado un modelo con baja varianza y alto sesgo. Por lo general, a medida que el modelo se vuelva más complejo (tenga más profundidad), se notará una disminución en el error de predicción debido a que se genera un menor sesgo.<br>* Un pequeño cambio en los datos puede cambiar drásticamente la estructura del árbol.<br>* Es muy común tener overfitting en el modelo.<br>* Si los datos que se quieren predecir tienen<br>información faltante, el árbol no generará las ramas correctas. |

| | | |
|---|---|---|
| **Random Forest** | * Se utilizan muy pocas suposiciones y, al igual que los árboles de decisión, la preparación de datos es mínima.<br>* Este modelo puede procesar gran cantidad de variables de entrada y seleccionar las más importantes.<br>* Funciona muy bien para problemas de clasificación.<br>* Tiene una forma efectiva de estimar los datos perdidos y mantener la precisión cuando se pierde una gran parte de los datos.<br>* Este método utiliza muchos árboles, por lo que reduce considerablemente el riesgo de overfitting. | * Al contrario de los árboles de decisión, la clasificación de este modelo es difícil de interpretar, dado que genera muchos árboles de decisión para llegar a la estimación.<br>* No funciona tan bien en problemas de regresiones.<br>* Poco control sobre el proceso de estimación del modelo, ya que es una caja negra para modeladores estadísticos, en comparación con los modelos anteriores (Logit y Árbol de decisión). |

Fuente: Elaboración propia, 2022.

En la siguiente tabla se presentan algunos de los exponentes de las metodologías de *credit score*:

**Tabla 3.** Estudios relacionados con el tema tratado

| Título de proyecto | Propósitos/Objetivos | Metodología | Modelos | Variables | Resultados |
|---|---|---|---|---|---|
| Perfil de riesgo de crédito para una cooperativa en Villavicencio, a partir de un modelo logit. (Pardo Carrillo, 2020) | Estimar el perfil de riesgos de los asociados (buenos y malos) de la cooperativa, que acceden a créditos en la modalidad de consumo. | Se plantea mínimo el 50% de deterioro de la cartera de consumo, mediante la metodología propuesta por Markov (Matrices de Transición). | *Logit | Edad, Ingreso Total, Plazo, Género Antigüedad, Monto Prestado, Empleado Sector Privado, Empleado Sector Oficial, Estado Civil y Personas a cargo. | El perfil de riesgos de la cooperativa, obedece a las siguientes características de los asociados: ser hombres, de poca edad, con menores ingresos, estar solteros, tener poca antigüedad como socios en la cooperativa, no tener personas a cargo, ser empleados de empresas privadas y tener créditos con montos altos, al igual de su respectivo plazo. |
| Predicción de riesgo crediticio en Colombia, usando técnicas de inteligencia artificial. (Borrero-Tigreros y Bedoya-Leiva, 2020) | Predicción de riesgo crediticio usando técnicas de inteligencia artificial, en una entidad financiera colombiana. | Utilización de tres tipos de algoritmos de clasificación supervisada. En la primera fase se consolida la base que contiene los datos básicos de los clientes con información de la operación crediticia y la información detallada de los pagos realizados en un periodo de tiempo, recopilados en 12 variables, | *Árboles de decisión aplicando el algoritmo C5.0 *Random Forest *Árboles de decisión aplicando el algoritmo C4.5 | Capital financiado, Género, Educación, Estado Civil, Edad, Historial de los últimos 6 meses de pago, Importe de estado de cuenta, Monto pagado de los últimos 6 meses, Cuotas pactadas, entre otras. | La estimación por Random Forest es más acertada que los demás modelos, puesto que dicho algoritmo obtuvo los valores más altos de especificidad (93.02%), exactitud (84.21%), precisión (72.72%), y área bajo la curva (88.29%), superando la sensibilidad obtenida por el modelo de redes neuronales (71.83%). |

| | | | | | |
|---|---|---|---|---|---|
| A note on random forest as the future credit scoring model. (Goh & Lee, 2020) | Incluir varias discusiones para descubrir el potencial de Random Forest como el futuro modelo de calificación crediticia. | Uso de dos conjuntos de datos de crédito (préstamos con garantía hipotecaria y préstamo entre pares de Lending Club a 60 meses) para demostrar la diferencia de rendimiento entre Logit y Random Forest (RF) | *Logit<br>*Random Forest | Monto Préstamo, Ingreso anual, endeudamiento, Cancelación dentro de los 12 meses, monto del pago, gravámenes fiscales, entre otras. | Para comparar el rendimiento de ambos modelos, se utilizaron los indicadores del área bajo la curva y el Accuracy, donde el RF obtuvo mejores resultados en comparación con el Logit en ambos conjuntos de datos, demostrando que, aunque es considerado por muchos como una caja negra por no tener una fórmula explícita, ha demostrado que tiene los mejores resultados de precisión. |
| Factores determinantes del incumplimiento en créditos de consumo de una cooperativa de ahorro y crédito (Tobar Vidal, Vergara Toro, & López Avilés, 2018) | Construcción de un modelo Logit para cuantificar el impacto y significancia de las variables, que explican el incumplimiento (default) y establecer una estrategia para aceptar o rechazar una solicitud de crédito. | Análisis de créditos de consumo, siendo necesario balancear la base a través de remuestreo de la clase de menor frecuencia de la muestra original, dada la desproporción entre socios buenos (sin default) y socios malos (con default). | *Regresión Logística binaria | 40 variables sociodemográficas y financieras del cliente, además de variables netamente del crédito. Agrupadas en 6 factores, los cuales son: crediticio, financiero, comportamiento crediticio, laboral, sociodemográficos y factores económicos. | Los resultados del modelo fueron satisfactorios en cuanto a la tasa de aciertos, errores y la capacidad predictiva. No obstante, es necesario que para una futura recalibración del mismo, se incorporen nuevas variables de comportamiento interno por tramos de mora, analizando diferentes escenarios de tiempo previos al momento del otorgamiento. |

Fuente: Elaboración propia, 2022.

## 5. Pruebas de bondad de ajuste

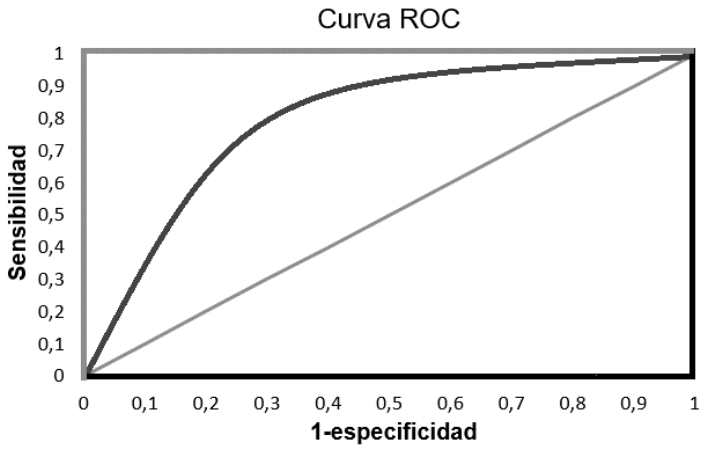
Para la elección del modelo más apropiado se utilizarán las siguientes pruebas diagnósticas:

### 5.1 Curva ROC

La curva ROC (Receiver Operating Characteristic, o Característica Operativa del Receptor) permite determinar la eficacia predictiva del modelo; gráficamente esta curva se construye con todos los pares entre sensibilidad (eje Y) y 1-especificidad (eje X), provenientes de la variación continua de los puntos de corte en todo el rango de los resultados.

Esta medida busca identificar un área bajo la curva y toma como referencia una diagonal que indica que el modelo es malo, puesto que no se logra diferenciar los clientes buenos sobre los malos; a medida que esta curva se vaya acercando a los extremos superiores del gráfico (ver figura 5), indica que el modelo presenta una mejor discriminación entre buenos y malos.

**Figura 5.** Curva de ROC




Fuente: Elaboración propia, 2022.

## 5.2 Test de Hosmer-Lemeshow

Es una prueba de bondad de ajuste, comúnmente usada en regresiones logísticas para comprobar si el modelo propuesto puede explicar lo observado; para este test la hipótesis nula es que el modelo es adecuado y la hipótesis alterna es que no lo es, por lo tanto, se busca no rechazar la hipótesis nula. El método que utiliza esta medida es agrupar las probabilidades de la variable dependiente, que predice el modelo y sacar las frecuencias esperadas para compararlas con las observadas mediante una prueba chi-cuadrado.

## 5.3 Matriz de confusión

Es una matriz de 2x2 que resume la clasificación del modelo frente a los datos observados, con el fin de medir el desempeño de la estimación.

**Tabla 4.** Matriz de confusión

| | | Predicción | |
|---|---|---|---|
| | | Cumplido = 0 | Incumplido = 1 |
| Real | Cumplido = 0 | Verdaderos Positivos | Falsos Negativos |
| | Incumplido = 1 | Falso Positivos | Verdaderos Negativos |

Fuente: Elaboración propia, 2022.

Los valores de la diagonal principal corresponden a los estimados de forma correcta por el modelo, es decir, tanto los verdaderos cumplidos como los verdaderos incumplidos, mientras que la diagonal secundaria representa los casos en los que el modelo se ha equivocado. A partir de la matriz de confusión se pueden calcular métricas de exactitud (Accuracy) y precisión (precisión); la primera determina el porcentaje de predicciones correctas frente al total de los casos, mientras que la

segunda hace referencia al cociente entre los verdaderos casos positivos y el total de predicciones positivas (verdaderos positivos + verdaderos negativos).

La métrica a elegir dependerá de las necesidades de la organización, toda vez que hay un costo de oportunidad al pretender tener más exactitud que precisión en el modelo, por lo que debe existir un balance. Adicionalmente, se cuenta con otras métricas como son: la sensibilidad o Recall, que indica la tasa de verdaderos positivos y especificidad de los verdaderos negativos.

## 5.4 Prueba de Kolmogorov-Smirnov o K-S

El estadístico de Kolmogorov–Smirnov, también denominado como distancia Kolmogorov–Smirnov (K-S), se describe como la distancia vertical máxima entre las funciones de distribución acumulada empíricas de dos muestras, o entre una función de distribución empírica y una función de distribución acumulada teórica de referencia. La ventaja principal de este estadístico radica en la sensibilidad a diferencias, tanto en la localización como en la forma de la función de distribución acumulada.

Para determinar el resultado de esta prueba se debe contrastar la hipótesis nula:

**$H_o$**: dos muestras independientes de tamaños n1 y n2 proceden de la misma población, es decir, las distribuciones son iguales ($p>0.05$)

**$H_1$**: dos muestras independientes de tamaños n1 y n2 no proceden de la misma población, es decir, las distribuciones son distintas ($p<0.05$)

## 6. Metodología

Para obtener el resultado esperado, se llevó a cabo una investigación de las metodologías que la literatura menciona, para realizar el perfilamiento de crédito a personas naturales, evaluando cuál era la más eficiente para la estimación del modelo.

Como la cooperativa no contaba con toda la información de los formularios de solicitud de crédito almacenada en el sistema contable en un reporte único, fue necesario realizar el cruce de información entre reportes diseñados, conforme a las directrices del ente regulatorio para transmisión y los demás reportes y listados que la organización albergaba en sus bases de datos.

La base de datos fue depurada teniendo en cuenta los siguientes criterios de exclusión:

- Asociados inactivos
- Asociados con datos faltantes o duplicados
- Asociados sin créditos vigentes al corte estudiado
- y demás consideraciones necesarias, así mismo, se realizó un análisis de cartera por edades para identificar el riesgo de crédito de las operaciones realizadas por la cooperativa (Anexo 1).

Una vez obtenida y depurada toda la información requerida, la base de datos para el desarrollo de este trabajo fue conformada únicamente por los asociados con créditos vigentes al año 2020. Posteriormente, se analizaron las diferentes metodologías que, de acuerdo con el tamaño de la muestra y la calidad de la información, podían ser usadas para finalmente estimar modelos tipo logit, árbol de decisión, Random Forest y modelo regulatorio, aunque este último por el momento no aplique para la Cooperativa.

Se realizaron pruebas estadísticas para determinar la significancia de los modelos y, una vez obtenidas las probabilidades de incumplimiento, se construyeron los rangos de calificación respectivos, para generar simulaciones con clientes nuevos y encontrar fortalezas y debilidades de cada modelo, logrando establecer cuál es el más eficiente para la organización.

# 7. Resultados

## 7.1 Desarrollo del modelo

Para el desarrollo del modelo se utilizó la base de datos suministrada por la Cooperativa, depurada bajo los criterios de inclusión y exclusión determinados en la metodología, considerando finalmente 273 asociados con créditos, en su mayoría catalogados como categoría de Riesgo A (96.06%), donde el 85.30% de la cartera pertenece a las líneas de crédito Libre inversión y Mejoramiento (Anexo 2).

Debido a la reducida cantidad de asociados clasificados en categoría de incumplimiento, para mejorar la precisión de la estimación, fue necesario balancear la base de datos, creando una nueva variable denominada: factor de ponderación, la cual toma el valor de 20[7] para el caso de los incumplidos y 1 para los no incumplidos, es decir, se realizó un remuestreo de la clase de menor frecuencia de la muestra original (oversampling[8]), con la proporción 66% de los asociados cumplidos y 34% de los incumplidos.

[7] Valor encontrado a través de simulaciones cuyo criterio de elección fue la significancia del modelo.
[8] Procedimiento mediante el cual se expande la clase minoritaria de una base de datos.

**Tabla 5.** Frecuencia por clase muestral

| Estado | Frecuencia | % | Factor ponderación | Frecuencia balanceada | % |
|---|---|---|---|---|---|
| Cumplido | 266 | 97% | 1 | 266 | 66% |
| Incumplido | 7 | 3% | 20 | 140 | 34% |
| **Total** | **273** | **100%** | | **406** | **100%** |

Fuente: Elaboración propia, 2022.

Adicionalmente, para identificar las variables independientes, se procedió a obtener la mayor información posible de los formatos de solicitud de crédito, bases de cartera y demás informes reportados a la Supersolidaria.

### 7.1.1 Variables

Se estableció que las variables que se iban a analizar en el modelo serían principalmente las reportadas por la cooperativa ante la Supersolidaria, eligiendo 12 en total, las cuales se listan a continuación:

1. **Género:** Sexo del asociado, toma el valor de cero si es hombre y uno si es mujer.
2. **Edad:** Edad en la cual se encuentra el asociado.
3. **Aportes:** Aportes sociales individuales que efectúa el asociado por una suma no inferior al dos por ciento (2%) ni superior al (10%) de su ingreso mensual.
4. **Codeudor:** 0 en caso de no tener créditos o no tener codeudor y 1 en caso de tener 1 o más codeudores.
5. **Estrato:** Estrato socioeconómico del prestatario que se ubica del 1 al 6. Cabe mencionar que la población objeto de estudio se encuentra concentrada en los estratos 2, 3 y 4 (Anexo 3).
6. **Estado Civil:** Soltero, casado, unión libre, separado, divorciado o viudo.
7. **Nivel Educativo:** Grado de escolaridad del asociado, que puede ser: Bachiller, Ninguno, Universitario, Postgrado, Primaria, Técnico, Tecnológico.
8. **Antigüedad:** Tiempo que lleva vinculado el asociado a la cooperativa.
9. **Nivel de Ingresos:** Agrupados en 7 niveles

- Menor a 1 SMMLV
- 1 a 2 SMMLV
- 2 a 3 SMMLV
- 3 a 4 SMMLV
- 4 a 6 SMMLV
- 6 a 8 SMMLV
- 8 a 11 SMMLV

10. **Valor Préstamo:** Monto total de capital, intereses y otros.
11. **Saldo Capital:** Monto de capital adeudado por el asociado.
12. **Porcentaje de Aportes:** % de aportes en relación al asalario del asociado.
13. **Cantidad Créditos:** Números de créditos otorgados al asociado.
14. **Mora Trimestral:** Mora presentada por el asociado en el último trimestre.
15. **Trabaja en la Universidad:** Toma el valor de 1 si no es empleado de la Universidad Autónoma, en caso contrario, es cero.

## 7.2 Análisis de datos en software

Para el desarrollo de los modelos se usaron los programas estadísticos: IBM SPSS Statistics 27.0, Stata versión 17 y Python versión 3.8, a partir de los cuales se puede realizar un análisis estadístico avanzado y trabajar con un lenguaje de programación con un código fácilmente legible de fácil comprensión y uso.

### 7.2.1 Modelo de regresión logística binaria

Para construir este modelo fue necesario determinar como variable dependiente, el incumplimiento en la obligación adquirida con la Cooperativa (obligaciones con mora posterior a los 30 días) y, como variables independientes, finalmente se establecieron: Trabaja en la Universidad, género, aportes, antigüedad, mora

trimestral, edad y nivel de ingresos. Por temas de calibración del modelo, se debieron incluir las variables Antigüedad y Edad transformadas en logaritmo y raíz, respectivamente, empleando la metodología de transformación de Box-Cox (Anexo 4).

Las anteriores consideraciones fueron aplicadas al modelo, arrojando las siguientes variables estadísticamente significativas, como se evidencia en la tabla 6:

**Ecuación 4**

$$Z = -345.4446 + 7.393572\ Trabaja\ U - 5.235967\ Género - 9.92e - 07\ Aportes - 1.426215\ Antigüedad + 12.0728\ MorTrim - 6.920893\ Edad - 9.52e - 07\ NiveldeIngresos + 95.99687\ Edad_Raíz + 30.90855\ Antigüedad_log$$

Donde:

**Ecuación 5**

$$P(Probabilidad\ de\ Incumplimiento) = \frac{1}{1 + e^{-(Z)}}$$

Las variables que intervienen en el modelo final se muestran a continuación:

**Tabla 6.** Variables en la ecuación-Modelo Logit

| | Coefficient | Sig |
|---|---|---|
| TrabajaU | 7.393572 | 0.000 |
| Género | - 5.235967 | 0.001 |
| Aportes_2020 | - 0.000001 | 0.000 |
| Antigüedad | - 1.426215 | 0.000 |
| MorTrim | 12.072800 | 0.000 |
| Edad_Cal_2020 | - 6.920893 | 0.003 |
| NiveldeIngresos | - 0.000001 | 0.010 |
| Edad_Raíz | 95.996870 | 0.003 |
| Antigüedad_log | 30.908550 | 0.000 |
| Constante | - 345.444600 | 0.002 |

Fuente: Datos arrojados por el software Stata, 2022.

Una vez obtenida la ecuación, se procede a analizar los coeficientes de las variables, evidenciando que los principales determinantes de la probabilidad de incumplimiento de los asociados son: el género, la edad y si trabaja o no en la Universidad; así mismo, se observa que haber incurrido en mora durante el último trimestre, supone un nivel de riesgo considerable al presentar un coeficiente alto.

Se tiene que a mayor antigüedad, menor probabilidad de incumplimiento, a mayor cantidad de aportes menor riesgo, dado que estos últimos hacen parte de las garantías constituidas para respaldar el crédito, lo cual indica la capacidad predictora de la ecuación arrojada por el modelo.

De acuerdo con la tabla 6, de los coeficientes de las nueve variables que componen el modelo, cuatro de ellas influyen de manera negativa y el resto de manera positiva en la probabilidad de que un asociado presente incumplimiento en el pago de sus obligaciones. Dado que se fija un $p - value$ en 0,05 y al contrastarse el nivel de significancia de las variables independientes, se concluye que estas influyen en el comportamiento de pago.

Para determinar la bondad de ajuste, se procede a realizar las pruebas mencionadas previamente en el marco teórico; una de ellas es la prueba de Hosmer-Lemeshow, dado su nivel de significancia P=0,9833>0,05 como se muestra en el Anexo 6; no se rechaza la hipótesis nula, lo que indica que el modelo se ajusta lo suficiente a los datos. Por otro lado, la prueba K-S indica que las distribuciones son iguales, dado su nivel de significancia P= 0,9932 >0,05.

Adicionalmente, se calculan las métricas de sensibilidad y especificidad para definir el punto de corte óptimo, que establece el umbral a partir del cual un asociado es catalogado como incumplido, siendo este el punto en el cual la distancia entre estas dos medidas sea la menor.

**Tabla 7.** Elección del punto de corte

| Punto Corte | Sensibilidad | Especificidad | Distancia |
|---|---|---|---|
| **50%** | 96,24% | 100,00% | -3,76% |
| **40%** | 95,49% | 100,00% | -4,51% |
| **30%** | 93,98% | 100,00% | -6,02% |
| **20%** | 91,35% | 100,00% | -8,65% |
| **10%** | 88,35% | 100,00% | -11,65% |

Fuente: Elaboración propia, 2022.

Las probabilidades de incumplimiento obtenidas en la regresión se comparan con el punto de corte óptimo, que para este ejercicio fue el 50% y con ello se procede a construir la matriz de confusión (tabla 8), para evaluar la capacidad predictiva del modelo, donde se observan los datos reales y los pronosticados por el mismo, distinguiendo entre los asociados que han presentado incumplimiento en el pago de los créditos y los que no.

**Tabla 8.** Matriz de confusión-Modelo Logit

| | | Predicción | | Total Casos | % |
|---|---|---|---|---|---|
| | | Cumplido = 0 | Incumplido = 1 | | |
| Real | Cumplido = 0 | 256 | 10 | 266 | 96,24% |
| | Incumplido = 1 | 0 | 7 | 7 | 0,00% |

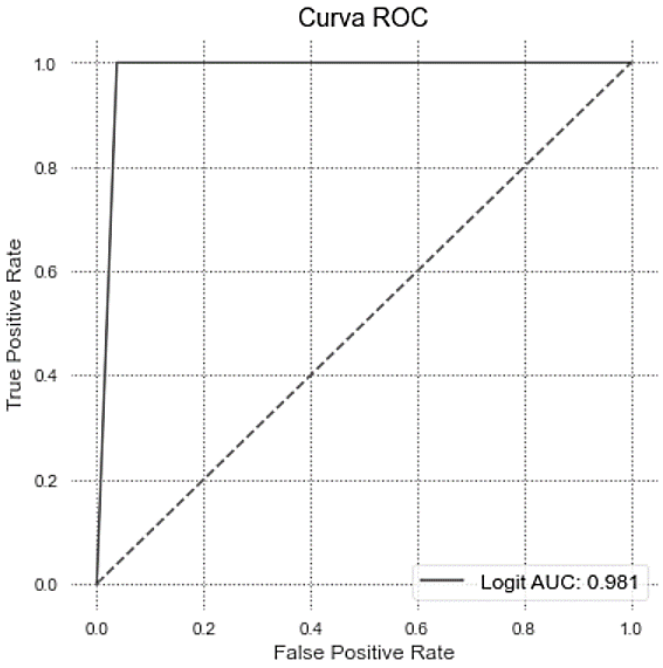
Fuente: Elaboración propia, 2022.

Como se observa en la tabla 8, el nivel de sensibilidad o acierto de verdaderos positivos en este modelo es del 96,24%, dado que de los 266 asociados cumplidos que se tenían en la base original, únicamente 256 fueron catalogados como cumplidos por el modelo. Por otro lado, se tiene un nivel de especificidad o acierto de verdaderos negativos del 100%, puesto que los siete asociados incumplidos que se tenían inicialmente siguen siendo catalogados como tal.

Por otra parte, el Accuracy (nivel de exactitud) es de 96,34%, cifra que es favorable, dado que esta medida es considerada como la capacidad global del modelo de predecir, tanto los verdaderos positivos como los verdaderos negativos, con respecto al total de las observaciones.

Finalmente, al observar la figura 6, es claro que la curva del modelo está cerca de la esquina superior izquierda, y el área bajo la curva es 0,981, lo que confirma que se tiene un buen poder discriminativo para clasificar los asociados cumplidos e incumplidos.

**Figura 6.** Curva de ROC-Modelo Logit




Fuente: Elaboración propia, 2022.

### 7.2.2 Modelo árboles de decisión

A partir de la base de datos depurada se corrió este modelo, utilizando el software SPSS con la metodología CHAID (chi-square automatic interaction detection), la cual elige la variable independiente (predictora) que presenta la interacción más fuerte con la variable dependiente (incumplimiento) y las categorías de cada predictor se descartan, si no son significativamente distintas respecto a la variable dependiente; en este caso, al tratarse de una base de datos pequeña, se eligió una configuración de 50/50 para los nodos padres e hijos respectivamente, lo que señala que cada nodo padre deberá tener como mínimo 50 casos y cada nodo hijo 50 (el software viene predeterminado con la configuración 100/50).

El árbol está conformado por ocho nodos y tres niveles de profundidad. En el primer nivel se encuentra el default (variable dependiente), se obtienen 5 perfiles de riesgo (tabla 9), los cuales tienen asociada una probabilidad de impago de acuerdo con las características que posee el asociado.

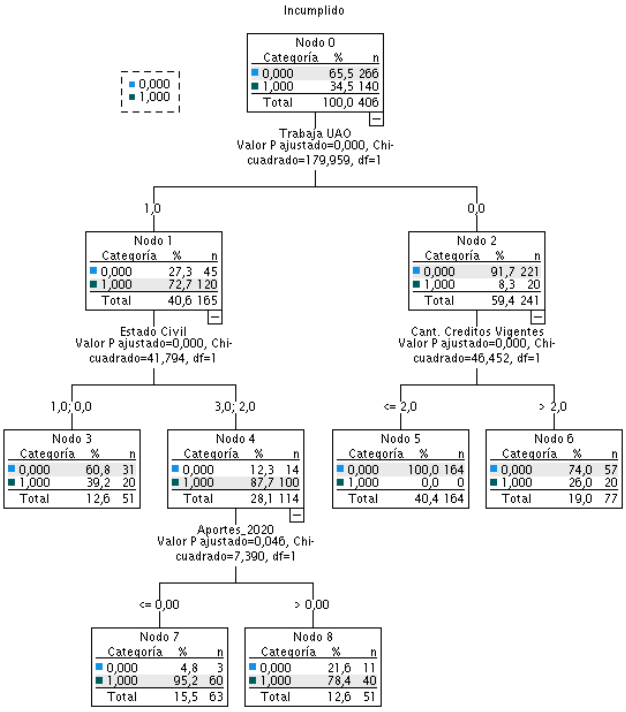
**Tabla 9.** Perfiles de riesgo

| Categoría | % de Incumplimiento |
|---|---|
| Categoría 1 | 95,2% |
| Categoría 2 | 78,4% |
| Categoría 3 | 39,2% |
| Categoría 4 | 25,8% |
| Categoría 5 | 00,0% |

Fuente: Elaboración propia, 2022.

A continuación, se despliegan las variables independientes resultantes más fuertes, es decir, las que tienen una mayor relevancia en el modelo: Trabaja en la universidad, estado civil, créditos vigentes y aportes.

**Figura 7.** Diagrama modelo árbol de decisión




Fuente: Datos arrojados por el software SPSS[9], 2022.

**Categoría 1:** Esta categoría muestra el perfil más riesgoso, pues tiene un nivel de incumplimiento de 95,2% al tratarse de asociados con aportes 0 pesos, estado civil Unión Libre (2) o Separado (3) y retirado de la universidad.

**Categoría 2:** Para que el Asociado se encuentre en esta clasificación debe tener un aporte mayor a 0 pesos, el estado civil debe ser Unión Libre (2) o Separado (3), y ser un ex – empleado de la universidad. Esta categoría tiene una probabilidad de

[9] Estado civil = 0: Soltero, 1: Casado, 2: Unión Libre y 3: Separado.

incumplimiento en el pago de sus obligaciones del 78,4%, por lo cual no existe la posibilidad de otorgar el crédito.

**Categoría 3:** Si el Asociado tiene un estado civil soltero (0) o casado (1) y además es ex – empleado de la universidad, se ubicará en esta categoría con una probabilidad de incumplimiento del pago de sus obligaciones del 39,2%. A pesar que es una probabilidad menor a la categoría anterior, es importante realizar un estudio más detallado para evaluar la posibilidad de otorgar el crédito.

**Categoría 4:** En esta categoría se observa que la probabilidad de incumplimiento en el pago de las obligaciones es del 25,8%. Se conforma por empleados de la universidad con más de dos créditos desembolsados.

**Categoría 5:** Si el Asociado tiene dos o menos créditos y trabaja con la universidad tiene una probabilidad de incumplimiento del 0%, es decir, posee un buen perfil para el otorgamiento.

Posteriormente, se procede a elegir el punto de corte óptimo; como se puede observar en la tabla 10, los puntos de corte 40% y 50% comparten los mismos niveles de sensibilidad y especificidad, al igual que 10% y 20% entre sí. No obstante, al evaluar la distancia entre la sensibilidad y la especificidad, se tiene que la menor en valor absoluto es la asociada al punto de corte 30% (-2.63%).

**Tabla 10.** Elección punto de Corte

| Punto Corte | Sensibilidad | Especificidad | Distancia |
|---|---|---|---|
| **50%** | 94,74% | 71,43% | 23,31% |
| **40%** | 94,74% | 71,43% | 23,31% |
| **30%** | 83,08% | 85,71% | -2,63% |
| **20%** | 61,65% | 100,00% | -38,35% |
| **10%** | 61,65% | 100,00% | -38,35% |

Fuente: Elaboración propia, 2022.

Las probabilidades de incumplimiento obtenidas en los perfiles de riesgo se comparan con el punto de corte elegido y con ello se procede a construir la matriz de confusión (tabla 11).

**Tabla 11.** Matriz de confusión- árboles de decisión

| | | Predicción | | Total Casos | % |
|---|---|---|---|---|---|
| | | Cumplido = 0 | Incumplido = 1 | | |
| Real | Cumplido = 0 | 221 | 45 | 266 | 83,08% |
| | Incumplido = 1 | 1 | 6 | 7 | 14,29% |

Fuente: Elaboración propia, 2022.

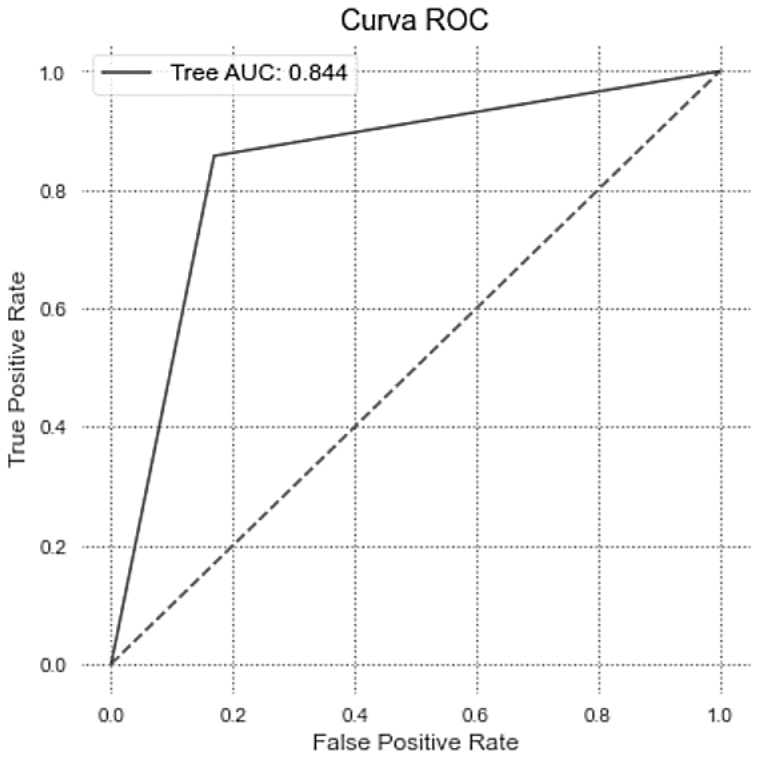
Como se observa en la tabla 11, el nivel de sensibilidad o acierto de verdaderos positivos en este modelo es del 83,08%, dado que de los 266 asociados cumplidos que se tenían en la base original, únicamente 221 fueron catalogados como cumplidos por el modelo. Por otro lado, se tiene un nivel de especificidad o acierto de verdaderos negativos del 85.71%, puesto que de los 7 asociados incumplidos que se tenían inicialmente, 6 continúan siendo catalogados como tal.

Hay que mencionar, que el Accuracy (nivel de exactitud) para este modelo es de 83,15%, indicando una buena clasificación de los asociados que verdaderamente son cumplidos e incumplidos.

Al observar la figura 8, es claro que la curva del modelo está cerca de la esquina superior izquierda, y el área bajo la curva es 0,844, lo que confirma que se tiene un buen poder discriminativo para clasificar los asociados cumplidos e incumplidos.

Para determinar la bondad de ajuste, se evalúa la prueba K-S, que indica que las distribuciones son diferentes, dado su nivel de significancia P= 0,001634 <0,05 (se rechaza $H_o$).

**Figura 8.** Curva ROC- árboles de decisión




Fuente: Elaboración propia, 2022.

### 7.2.3 Random Forest

Para este modelo de Maching learning se utilizó el entorno computacional interactivo Jupyter Notebook, ya que es compatible con el lenguaje de programación Python y para ejecutarlo se usaron las librerías de Pandas, Numpy y Scikit-learn.

Teniendo en cuenta que para correr el modelo se deben determinar los hiperparámetros más adecuados para el RandomForrestClassifier de Sklearn, se recurrió a la función ParameterGrid que analiza varias combinaciones de los mismos, que obtienen el Accuracy más alto posible.

De las combinaciones analizadas se obtuvieron 2 modelos de Random Forest que se adaptan mejor a la naturaleza del problema, estos utilizan 50 árboles y una profundidad máxima de 10. Sin embargo, para el primer modelo se utilizó una class_weight = {1: 40} y para el segundo fue de {1:10}, esto con el fin de balancear las clases en los datos (cumplidos e incumplidos) y que estos algoritmos clasifiquen mejor la clase minoritaria que, en este caso, son los incumplidos.

Dado que la información para la estimación de estos modelos utiliza variables categóricas, se realizó la técnica de one-hot encoding, la cual crea una columna para cada valor distinto de la variable categórica y para cada registro coloca 1 a la columna a la que pertenezca dicho registro y deja las demás con 0.

**Figura 9.** Proceso one-hot encoding

| Variable Categórica |
|---|
| 1 |
| 2 |
| 3 |
| 4 |
| 2 |
| 3 |
| 1 |

One-Hot encoding

| Variable Categórica_1 | Variable Categórica_2 | Variable Categórica_3 | Variable Categórica_4 |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 |

Fuente: Elaboración propia, 2022.

Con los anteriores parámetros se desarrollaron los dos modelos propuestos de Random Forest, no obstante, tal y como se realizó en los modelos anteriores (Logit y árbol de decisión), se procede a seleccionar el punto de corte óptimo (tabla 12), donde 5% es el valor elegido para ambos Random Forest, dada la distancia corta que hay entre la sensibilidad y especificidad, como ya se ha mencionado.

**Tabla 12**. Elección de puntos de corte – modelos de Random Forest

Random Forest 1

| Punto Corte | Sensibilidad | Especificidad | Distancia |
|---|---|---|---|
| **50%** | 99,62% | 85,71% | 13,91% |
| **40%** | 99,62% | 85,71% | 13,91% |
| **30%** | 98,50% | 85,71% | 12,78% |
| **20%** | 96,99% | 85,71% | 11,28% |
| **10%** | 99,60% | 92,86% | 6,74% |
| **5%** | 86,47% | 85,71% | 0,75% |

Random Forest 2

| Punto Corte | Sensibilidad | Especificidad | Distancia |
|---|---|---|---|
| **50%** | 100,00% | 85,71% | 14,29% |
| **40%** | 100,00% | 85,71% | 14,29% |
| **30%** | 100,00% | 85,71% | 14,29% |
| **20%** | 100,00% | 85,71% | 14,29% |
| **10%** | 98,87% | 85,71% | 13,16% |
| **5%** | 94,74% | 85,71% | 9,02% |

Fuente: Elaboración propia, 2022.

Las probabilidades de incumplimiento obtenidas bajo esta técnica de aprendizaje automático, se contrastan con el punto de corte óptimo (5%), para obtener la matriz de confusión (tabla 13), y así valorar la capacidad predictiva de estos modelos.

**Tabla 13.** Matriz de confusión- modelos de Random Forest

Random Forest 1

| | | Predicción | | Total Casos | % |
|---|---|---|---|---|---|
| | | Cumplido = 0 | Incumplido = 1 | | |
| Real | Cumplido = 0 | 230 | 36 | 266 | 86,47% |
| | Incumplido = 1 | 1 | 6 | 7 | 14,29% |

Random Forest 2

| | | Predicción | | Total Casos | % |
|---|---|---|---|---|---|
| | | Cumplido = 0 | Incumplido = 1 | | |
| Real | Cumplido = 0 | 252 | 14 | 266 | 94,74% |
| | Incumplido = 1 | 1 | 6 | 7 | 14,29% |

Fuente: Elaboración propia, 2022.

Como se puede inferir en la tabla 9, el nivel de sensibilidad en estos modelos es de 86,47% y 94,74%, dado que de los 266 asociados cumplidos que tiene la cooperativa en el periodo de estudio, 230 y 252 fueron catalogados como cumplidos por los modelos respectivamente. Así mismo, el nivel de especificidad es de 85,71% para ambos, debido a que, de los 7 asociados incumplidos reales, el modelo clasifica bajo esta categoría a 6.

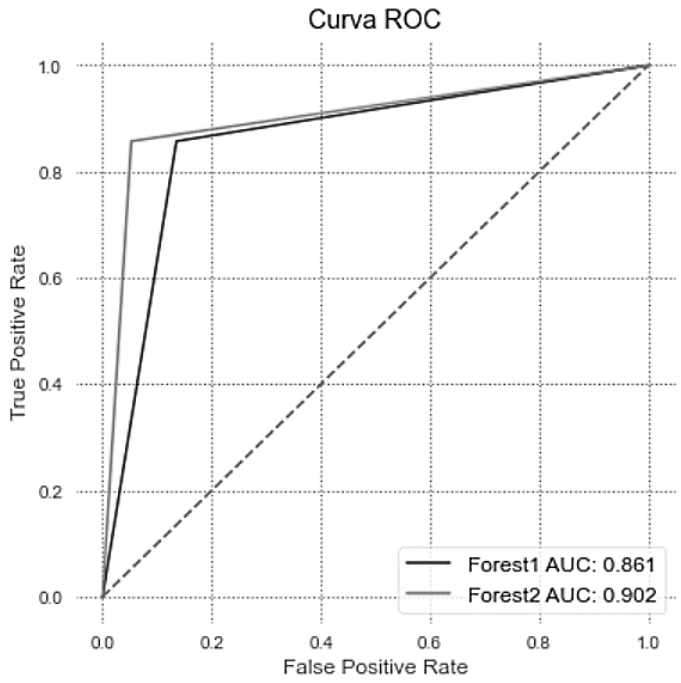
Adicionalmente, el Accuracy (nivel de exactitud) es de 86,45% y 94,51%, lo que demuestra que el segundo Random Forest tiene una mejor capacidad de predecir para clasificar los asociados cumplidos e incumplidos.

Para determinar la bondad de ajuste, se evalúa la prueba K-S, que indica que las distribuciones son diferentes para el modelo Random Forest 1, dado su nivel de

significancia P= 0,022412 <0,05 (se rechaza Ho), mientras que para el modelo Random Forest 2 son iguales (P= 0,917071953<0,05).

La figura 10 muestra el área bajo la curva de ambos modelos 0,861 y 0,902, confirmando que el segundo modelo tiene un mejor poder discriminativo para clasificar los asociados cumplidos e incumplidos que el primero.

**Figura 10.** Curva ROC-Modelos Random Forest




Fuente: Elaboración propia, 2022.

Los modelos Random Forest 1 y 2 coinciden en que las variables más importantes para estimar la probabilidad de incumplimiento son: Trabaja en la universidad, los aportes, porcentaje de aportes, la edad, el nivel de ingresos y la antigüedad por mencionar algunas (Anexos 7 y 8).

## 7.3 Comparación entre los modelos

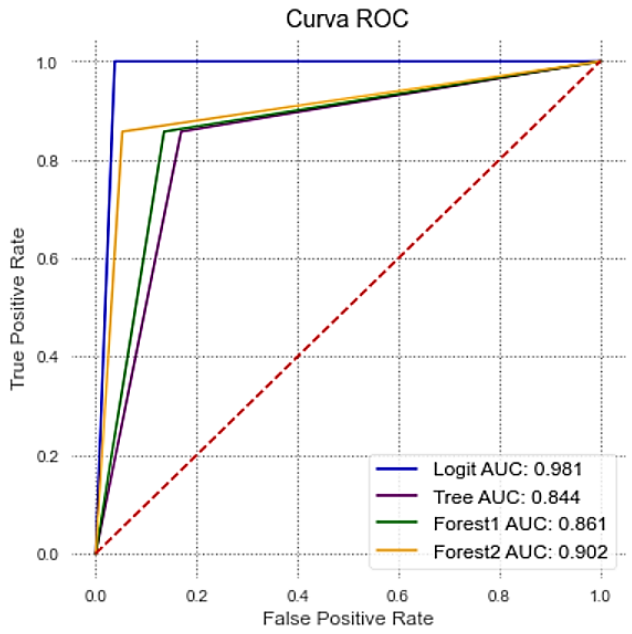
Para finalizar, se sintetiza la validación de los modelos desarrollados por medio de las tres metodologías, tal como se muestra en la tabla 14, comparando entre ellos el nivel de especificidad, sensibilidad, porcentaje global (Accuracy) y AUC.

**Tabla 14.** Comparación entre modelos

| | Modelo Logit | Árbol de decisión | Random Forest 1 | Random Forest 2 |
|---|---|---|---|---|
| Especificidad | 100,00% | 85,71% | 85,71% | 85,71% |
| Sensibilidad | 96,24% | 83,08% | 86,47% | 94,74% |
| AUC | 98,1% | 84,4% | 86,1% | 90,2% |
| K-S | 0,993233 | 0,001635 | 0,022413 | 0,917072 |
| Accuracy | 96,34% | 83,15% | 86,45% | 94,51% |

Fuente: Elaboración propia, 2022.

**Figura 11.** Curva ROC-Comparación Modelos




Fuente: Elaboración propia, 2022.

Con base en los puntajes arrojados por cada uno de los modelos estimados y los determinados por el modelo regulatorio para la modalidad consumo con libranza, se determinó la calificación de cada asociado según la escala establecida, como se detalla a continuación:

**Tabla 15.** Rangos de calificación por modelo

| Calificación | Regulatorio (Consumo con Libranza) | Logit | Árbol Decisión | Random Forest 1 | Random Forest 2 |
|---|---|---|---|---|---|
| A | 1.74% | 12.50% | 9.33% | 1.25% | 1.25% |
| B | 3.37% | 25.00% | 16.22% | 2.50% | 2.50% |
| C | 4.79% | 37.50% | 23.11% | 3.75% | 3.75% |
| D | 8.12% | 50.00% | 30.00% | 5.00% | 5.00% |
| E | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |

Fuente: Elaboración propia, 2022.

Al comparar los rangos de calificación de los modelos estimados versus el regulatorio, vemos que los más cercanos a este último en casi todas sus categorías son los Random Forest, que con un rango de hasta 1.25% califican al asociado en A, mientras que el regulatorio lo realiza con 1.74%; con un rango superior a 1.25% y hasta 2.50% califican al asociado en B, el cual también es relativamente cercano al modelo regulatorio con un límite de 3.37%. Por su parte, los modelos Logit y Árbol de decisión son menos ácidos que el regulatorio y los modelos Random Forest, puesto que tienen rangos más amplios para clasificar los asociados en cada categoría de riesgo.

Una vez construidos los rangos de calificación, se integraron a las calculadoras diseñadas para estimar la calificación que debe otorgarse al cliente, según las condiciones y respectivas variables de cada modelo estimado (Anexo 9). Dichas calculadoras fueron realizadas en Excel, empleando los coeficientes de las variables significativas para el caso del Logit y las categorías construidas para el caso del

árbol de decisión; esta herramienta permite ingresar las características del cliente (variables elegidas en cada modelo) y obtener una probabilidad de incumplimiento, que se traduce en una calificación de acuerdo a los rangos establecidos para cada modelo.

Las calificaciones obtenidas son una estimación que permite a la entidad continuar con el desarrollo de modelos propios que son útiles en la toma de decisiones, no obstante, actualmente realiza la constitución de la pérdida esperada por altura de mora con factor de riesgo, es decir, la calificación es determinada por los días de mora en que haya incurrido el asociado al momento de la evaluación y, de este modo, se obtiene la probabilidad de incumplimiento del mismo:

**Tabla 16.** Cálculo del deterioro para la modalidad de consumo

| Calificación | Consumo | Rango Días |
|---|---|---|
| A | 0% | Hasta 30 |
| B | 1% | más de 30 hasta 60 |
| C | 10% | más de 60 hasta 90 |
| D | 20% | más de 90 hasta 180 |
| E | 50% | más de 180 |
| Incumplimiento | 100% | |

Fuente: Anexo 1, Consideraciones para la calificación y deterioro de la cartera de crédito, Capitulo II Sistema de Administración del Riesgo de Crédito, Circular Básica Contable y Financiera, 2020.

Conviene subrayar que según las disposiciones de la Supersolidaria (2020), las organizaciones solidarias deberán calificar a los deudores en categorías de mayor riesgo, cuando producto del análisis del análisis cualitativo y cuantitativo, se evidencien riesgos adicionales que sustenten dicho cambio.

## 8. Conclusiones y recomendaciones

En el desarrollo de este trabajo se diseñó un modelo de *credit scoring* para la cartera de la Cooperativa Cooautónoma, entidad que ofrece servicios y beneficios a egresados, directivos, profesores y demás trabajadores de la Universidad Autónoma De Occidente, la cual está regulada por la Supersolidaria. Se identificó un cambio en la normatividad con relación al modelo de referencia para el cálculo de la pérdida esperada de las organizaciones pertenecientes al sector solidario, sin embargo, la cooperativa no debe acogerse a las nuevas disposiciones al no cumplir con todos los requerimientos y continuará constituyendo el deterioro por altura de mora con factor de riesgo, por lo que el desarrollo del modelo propuesto puede ser útil ante posibles cambios en normatividad y en la actualidad contribuir a que la evaluación del perfil crediticio de los asociados se realice de una forma menos subjetiva, contribuyendo a la mitigación del riesgo de crédito.

Para el análisis de la información se llevó a cabo el desarrollo de cuatro modelos, uno mediante regresión logística binaria, otro con árbol de decisión y dos a través de la técnica de Random Forest; todos presentaban un Accuracy superior al 80% e incluían entre las variables independientes: Trabaja en la universidad, género, aportes, nivel de ingresos, edad, antigüedad, entre otras. Sin embargo, la técnica de regresión logística binaria cuenta con el Accuracy más alto (96,34%); además de integrar variables significativas que guardan una relación coherente con la variable dependiente.

Otro indicador considerado para la elección del modelo es el área bajo la curva ROC, dada su utilidad para evaluar la capacidad discriminativa, es decir, diferenciar entre asociados cumplidos e incumplidos; siendo de 98,1% en la regresión logística binaria versus los demás modelos que se encuentran por debajo del 91%, se puede inferir que el modelo más adecuado para la organización es el Logit seguido por el Random Forest. Mientras que en los estudios realizados en determinadas entidades financieras colombianas, el modelo que reporta mayor precisión es el Random Forest.

De todos los modelos el que reporta un desempeño inferior a los demás fue el árbol de decisión, con un Accuracy del 83,15%, además de las pocas variables que utiliza para perfilar al asociado (trabaja universidad, estado civil, cantidad de créditos vigentes y aportes), dejando por fuera muchas combinaciones posibles para la clasificación de riesgo del asociado, lo cual puede dificultar la toma de decisiones acertadas.

Conviene subrayar que la idoneidad de los modelos estará definida no solo por las bondades estadísticas de los mismos, sino también por lo alineados que estén con el apetito de riesgo que desee asumir la entidad y acorde con los objetivos del negocio.

Se recomienda a la entidad realizar seguimiento a la captura de la información de los formularios de solicitud de crédito y considerar la inclusión de algunas variables adicionales a las sugeridas por el ente regulatorio en los reportes transmitidos, las cuales podrían generar valor agregado para el desarrollo de futuros modelos de riesgo, garantizando la calidad y significancia de los mismos.

Teniendo en cuenta las limitaciones en la recopilación de la información, en temas de: cantidad de datos, variables y número de clientes incumplidos, hubo la necesidad de realizar ponderaciones para equilibrar la base y transformaciones a ciertas variables, buscando mejorar la precisión de los resultados obtenidos. Por lo cual, se recomienda acompañar las estimaciones del modelo con la opinión de las áreas que intervienen en el proceso de crédito.

Es importante mencionar que, dadas las particularidades del modelo, se sugiere a la cooperativa actualizarlo mínimo dos veces al año, para monitorear su desempeño y evaluar correctamente el potencial deterioro de la calidad crediticia de los asociados.

## Referencias


Acero Sánchez, C. E. (2019). *Cooperativas: modelo para el desarrollo social.* Obtenido de Portafolio: https://www.portafolio.co/economia/cooperativas-modelo-para-el-desarrollo-social-525358

Alberca, A. (s.f.). *La librería Numpy*. Obtenido de Aprende con Alf: https://aprendeconalf.es/docencia/python/manual/numpy/

Alberca, A. (s.f.). *La librería Pandas*. Obtenido de Aprende con Alf: https://aprendeconalf.es/docencia/python/manual/pandas/

Alvear, J. (s.f.). *Árboles de decisión y Random Forest*. Obtenido de Bookdown.org: https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html

Amat Rodrigo, Joaquín (2020). *Machine Learning con Python y Scikitlearn.* Obtenido de Cienciadedatos.net: https://www.cienciadedatos.net/documentos/py06_machine_learning_python_scikitlearn.html

Borrero Tigreros, D., & Bedoya Leiva, O. (2020). Predicción de riesgo crediticio en Colombia usando técnicas de inteligencia artificial. *Revista UIS Ingenierías*, 37-52.

Breiman, L. (2001). Random Forest. *Maching Learning*, 5-32. Obtenido de https://doi.org/10.1023/A:1010933404324

Burgueño, M., García-Bastos, J., & González-Buitrago, J. (1995). Las curvas ROC en la evaluación de las pruebas diagnósticas. *Medicina clinica*, 661-670.

Cerda, J., & Cifuentes, L. (2012). Uso de curvas ROC en investigación clínica. Aspectos teórico-prácticos. *Revista Chilena de Infectología*, 138-141.

Cramer, J. (2003). *Logit Models from Economics and other Fields*. https://www.cambridge.org/core/books/logit-models-from-economics-and-other-fields/510D1D6FF8E2B6D6C80FEF7B4697EAB7

Goh, R., & Lee, L. (2020). A note on random forest as the future credit scoring model. *Journal of Analysis and Applications*, 105-117.

Gutiérrez Girault, M. (2007). Modelos de Credit Scoring -Qué, Cómo, Cuándo y Para Qué-. *MPRA Paper*.

Hand, D., & Henley, W. (1997). Stadistical Classification Methods in Consumer Credit. *J.R. Statist. Soc. A*, 523-541.

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recornition*, 278-282. Obtenido de https://doi.org/10.1109/ICDAR.1995.598994

ICHI PRO (s.f.). *¿Qué es la función sigmoidea? ¿Cómo se implementa en regresión logística?* Recuperado de https://ichi.pro/es/que-es-la-funcion-sigmoidea-como-se-implementa-en-regresion-logistica-77981969140323

Llopis Pérez, J. (s.f.). *La estadística: Una orquesta hecha instrumento*. Obtenido de Curso de estadistica: https://jllopisperez.com/2013/12/19/test-de-hosmer-y-lemeshow/

Millan Solarte, J. C., & Caicedo Cerezo, E. (2018). Modelos para otorgamiento y seguimiento en la gestión de riesgo de crédito. *Revista de Métodos Cuantitavios para la Economía y la Empresa*, 23-41.

Mures Quintana, J., García Gallego, A., & Vallejo Pascual, E. (2005). Aplicación del Análisis Discriminante y Regresión Logística en el estudio de la morosidad en las entidades financieras. Comparación de resultados. *Pecvnia*, 175-199.

Mylonakis, J., & Diacogiannis, G. (2010). Evaluating the Likelihood of Using Linear Discriminant Analysis as A Commercial Bank Card Owners Credit Scoring Model. *International Business Research*, 11.

Pardo Carrillo, O. S. (2020). Perfil de riesgo de crédito para una cooperativa en Villavicencio a partir de un modelo logit. *Universidad & Empresa, 22(38)*, 237-256.

Pérez López, C., & Santín González, D. (2008). *Minería de datos: Técnicas y herramienta.* Madrid: Paraninfo.

Rayo Cantón, S., Lara Rubio, J., & Camino Blasco, D. (2010). *Un Modelo de Credit Scoring para instituciones de microfinanzas en el marco de Basilea II.* Obtenido de Journal of Economics, Finance and Administrative: http://www.scielo.org.pe/scielo.php?pid=S2077-18862010000100005&script=sci_arttext&tlng=en

Rodríguez Caballero, C. V., & Espin García, O. (2013). Metodología para un scoring de clientes sin referencias crediticias. *Cuadernos de Economía*, 139-165.

SaludMadrid. (s.f.). *Modelo múltiple.* Obtenido de Hospital Universitario Ramon y Cajal: http://www.hrc.es/bioest/Reglog_5.html

Soporte de minitab 20 (s.f.). *¿Qué es un procedimiento bootstrap?* Obtenido de ¿Qué es un procedimiento bootstrap?: https://support.minitab.com/es-mx/minitab/20/help-and-how-to/probability-distributions-random-data-and-

resampling-analyses/supporting-topics/resampling-analyses/what-is-bootstrapping/

Stata (Nº de versión 17). (2022). StataCorp.

SPSS (N° de versión 27). (2022). IBM SPSS Statistics.

Supersolidaria. (2020). *Circular Básica Contable y Financiera (circular externa 22 de 2020).* Bogotá: Superintendencia de la Economía Solidaria. Obtenido de http://supersolidaria.gov.co/sites/default/files/public/documento_tecnico_circular_basica_contable_y_financiera_2020.pdf

TIBCO Software (s.f.). *What is a Random Forest?* TIBCO Software. Recuperado de https://www.tibco.com/reference-center/what-is-a-random-forest

Tobar Vidal, A., Vergara Toro, R., & López Avilés, D. (2018). Factores Determinantes del Incumplimiento en créditos de consumo de una cooperativa de ahorro y crédito. *Revista Chilena de Economía y Sociedad*, 40-60.

Yiu, T. (6 de 12 de 2019). *Understanding random forest*. Obtenido de Towards Data Science: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

## Anexos

**Anexo 1.** Cartera por edades a 31 de diciembre de 2020 (expresada en pesos)

| Edades | Valor Préstamos | Porcentaje |
|---|---|---|
| 0 a 30 días | $ 2.805.877.444,00 | 96,42% |
| 31 a 60 días | $ 9.049.000,00 | 0,31% |
| 61 a 90 días | $ 17.507.000,00 | 0,60% |
| 91 a 120 días | $ 26.885.000,00 | 0,92% |
| 121 a 150 días | $ - | 0,00% |
| Más de 150 días | $ 50.884.718,00 | 1,75% |
| **Cartera Total** | **$ 2.910.203.162,00** | **100,00%** |

Fuente: Elaboración propia con datos internos de la entidad, 2022.

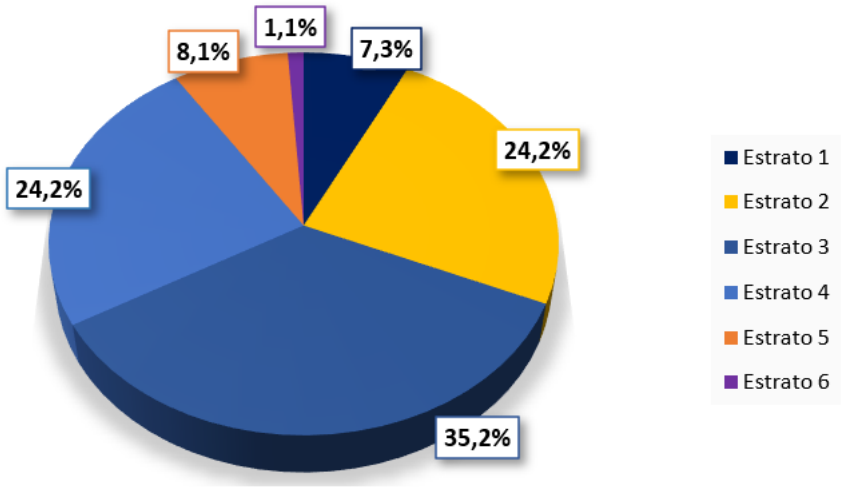
**Anexo 2.** Cartera por línea de crédito y categoría de riesgo a 31 de diciembre de 2020 (expresada en pesos)

| Líneas de Crédito | Valor Préstamos | Porcentaje |
|---|---|---|
| CREDIYA | $ 3.047.803,00 | 0,10% |
| EDUCACION | $ 51.588.091,00 | 1,77% |
| GASOLINA | $ 132.097,00 | 0,00% |
| LIB.INVERSIÓN | $ 2.240.858.554,00 | 77,00% |
| MEJORAMIENTO | $ 241.517.703,00 | 8,30% |
| MEMBRECÍA | $ 810.000,00 | 0,03% |
| PREFERENCIAL | $ 216.577.883,00 | 7,44% |
| REESTRUCTURAC | $ 29.670.613,00 | 1,02% |
| SEG. VEHÍCULO | $ 13.933.662,00 | 0,48% |
| TEM.ESPECIALE | $ 39.844.072,00 | 1,37% |
| VEHÍCULO | $ 72.222.684,00 | 2,48% |
| **Cartera Total** | **$ 2.910.203.162,00** | **100,00%** |

| Categoría Riesgo | Valor Préstamos | Porcentaje |
|---|---|---|
| CATEGORÍA A RIESGO NORMAL | $ 2.795.577.444,00 | 96,06% |
| CATEGORÍA B RIESGO ACEPTABLE | $ 9.049.000,00 | 0,31% |
| CATEGORÍA C RIESGO APRECIABLE | $ 17.807.000,00 | 0,61% |
| CATEGORÍA D RIESGO SIGNIFICATIVO | $ 28.885.000,00 | 0,99% |
| CATEGORÍA E RIESGO DE INCOBRABILIDAD | $ 58.884.718,00 | 2,02% |
| **Cartera Total** | **$ 2.910.203.162,00** | **100,00%** |

Fuente: Elaboración propia con datos internos de la entidad, 2022.

**Anexo 3.** Base de cliente por estratos




Fuente: Elaboración propia con datos internos de la entidad, 2022.

**Anexo 4.** Transformación de Box-Cox

Esta metodología viene dada para diferentes valores de λ por la siguiente expresión:

**Ecuación 6**

$$\left\{\begin{matrix} \dfrac{x^{\lambda}-1}{\lambda} & si\ \ \lambda = 0 \\ \log(x) & si\ \ \lambda \neq 0 \end{matrix}\right\}$$

Siendo Y la variable a ser transformada y λ el parámetro de transformación. A continuación, las transformaciones más habituales son:

| λ | Transformación |
|---|---|
| -2 | $\frac{1}{x^2}$ |
| -1 | $\frac{1}{x}$ |
| -0.5 | $\frac{1}{\sqrt{x}}$ |
| 0 | log (X) |
| 0.5 | $\sqrt{x}$ |
| 1 | $x$ |
| 2 | $x^2$ |

**Anexo 5.** Tabla de contingencia para la prueba de Hosmer y Lemeshow-modelo Logit

| Group | Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
|---|---|---|---|---|---|---|
| 1 | 0.0000 | 0 | 0.0 | 28 | 28.0 | 28 |
| 2 | 0.0000 | 0 | 0.0 | 27 | 27.0 | 27 |
| 3 | 0.0000 | 0 | 0.0 | 27 | 27.0 | 27 |
| 4 | 0.0000 | 0 | 0.0 | 28 | 28.0 | 28 |
| 5 | 0.0000 | 0 | 0.0 | 27 | 27.0 | 27 |
| 6 | 0.0001 | 0 | 0.0 | 27 | 27.0 | 27 |
| 7 | 0.0013 | 0 | 0.0 | 28 | 28.0 | 28 |
| 8 | 0.0125 | 0 | 0.1 | 27 | 26.9 | 27 |
| 9 | 0.2568 | 0 | 2.4 | 27 | 24.6 | 27 |
| 10 | 0.7341 | 20 | 19.2 | 12 | 12.8 | 32 |

Fuente: Datos arrojados por el software Stata, 2022.

**Anexo 6.** Prueba de Hosmer y Lemeshow-modelo Logit

| Number of observations | Number of groups | Hosmer–Lemeshow chi2 | Prob > chi2 |
|---|---|---|---|
| 273 | 10 | 2.92 | 0.9395 |

Fuente: Datos arrojados por el software Stata, 2022.

**Anexo 7.** Variables significativas modelo Random Forest 1

| | |
|---|---|
| Trabaja UAO | 0.152982 |
| Aportes_2020 | 0.132604 |
| Vlr_prest | 0.124714 |
| porc_aportes | 0.120821 |
| Antiguedad | 0.118718 |
| EDAD_cal_2020 | 0.073745 |
| nivel_ing_2 | 0.051764 |
| cant_cred | 0.044452 |
| estado_civil_UNION LIBRE | 0.031352 |
| estado_civil_SOLTERO (A) | 0.031291 |
| GENERO | 0.019252 |
| Escolar_2 | 0.018735 |
| codeudor | 0.009970 |
| Estrato_3 | 0.008724 |
| Estrato_2 | 0.008645 |
| Escolar_4 | 0.008467 |
| Estrato_5 | 0.007620 |
| nivel_ing_5 | 0.007398 |
| nivel_ing_3 | 0.006887 |
| Estrato_4 | 0.006783 |
| Escolar_5 | 0.005585 |
| Escolar_6 | 0.003557 |
| Escolar_3 | 0.002304 |
| nivel_ing_4 | 0.002142 |
| estado_civil_SEPARADO (A) | 0.001259 |
| Escolar_7 | 0.000229 |
| estado_civil_VIUDO (A) | 0.000000 |
| nivel_ing_6 | 0.000000 |
| Estrato_6 | 0.000000 |
| Escolar_8 | 0.000000 |
| Escolar_9 | 0.000000 |
| nivel_ing_7 | 0.000000 |

dtype: float64

Fuente: Datos arrojados por el software Python, 2022.

**Anexo 8.** Variables significativas modelo Random Forest 2

| | |
|---|---|
| Aportes_2020 | 0.147196 |
| Trabaja UAO | 0.143208 |
| porc_aportes | 0.127232 |
| Vlr_prest | 0.103769 |
| Antiguedad | 0.089639 |
| EDAD_cal_2020 | 0.069580 |
| nivel_ing_2 | 0.040693 |
| estado_civil_UNION LIBRE | 0.034005 |
| cant_cred | 0.029762 |
| Escolar_7 | 0.025170 |
| Escolar_2 | 0.023100 |
| Escolar_5 | 0.018082 |
| codeudor | 0.018010 |
| GENERO | 0.017711 |
| nivel_ing_5 | 0.014979 |
| estado_civil_SEPARADO (A) | 0.013835 |
| estado_civil_SOLTERO (A) | 0.013575 |
| Estrato_4 | 0.012369 |
| Estrato_2 | 0.011782 |
| Estrato_5 | 0.010645 |
| nivel_ing_3 | 0.009898 |
| Escolar_4 | 0.009787 |
| Estrato_3 | 0.009436 |
| Escolar_3 | 0.002011 |
| nivel_ing_4 | 0.001641 |
| nivel_ing_6 | 0.001333 |
| Escolar_9 | 0.001141 |
| Escolar_6 | 0.000411 |
| estado_civil_VIUDO (A) | 0.000000 |
| Estrato_6 | 0.000000 |
| Escolar_8 | 0.000000 |
| nivel_ing_7 | 0.000000 |

dtype: float64

Fuente: Datos arrojados por el software Python, 2022.

**Anexo 9.** Calculadoras rangos de calificación

## ESTIMACIÓN RANGO CRÉDITO CON MODELO LOGIT

| | | |
|---|---|---|
| **No. Identificación:** | xxx | |
| **Nombre:** | xyz | |
| **Fecha:** | 12/03/2022 | |
| **Capacidad de Pago Utilizada** | 24.85% | |
| **% Descuento** | 18.31% | **No excede el 50%** |
| **Saldo Cartera Total (K+I+O)** | 6,072,762 | |

| Variable | Dato | Betas |
|---|---|---|
| Constante | 1 | - 345.4446 |
| TrabajaUAO | 0 | 7.3936 |
| Ingreso_Group | 1 | 1.8010 |
| Género | 1 | - 5.2360 |
| Edad | 0 | - 6.9209 |
| Nivel de Ingresos | 1 | - 0.0000 |
| Antigüedad | 0 | - 1.4262 |
| MorTrim | 1 | 12.0728 |
| Edad Raíz | 0 | 95.9969 |
| Antigüedad_log | 1 | 30.9086 |
| AP | 1 | - 0.0000 |
| Z | -305.8982189 | |
| Puntaje | 1.4128E-133 | |
| Calificación | A | |

**Rangos Calificación estimación Modelo Logit**

| Calificación | Desde | Hasta | Calificación |
|---|---|---|---|
| A | 0.000% | 12.500% | A |
| B | 12.500% | 25.000% | B |
| C | 25.000% | 37.500% | C |
| D | 37.500% | 50.000% | D |
| E | 50.000% | 100.000% | E |

Fuente: Elaboración propia, 2022.

## ESTIMACIÓN RANGO CRÉDITO CON MODELO ÁRBOL DE DECISIÓN

| | | |
|---|---|---|
| **No. Identificación:** | xxx | |
| **Nombre:** | xyz | |
| **Fecha:** | 12/03/2022 | |
| **Capacidad de Pago Utilizada** | 24.85% | |
| **% Descuento** | 18.31% | **No excede el 50%** |
| **Saldo Cartera Total (K+I+O)** | 6,072,762 | |

| Variable | Dato |
|---|---|
| TrabajaUAO | 1 |
| Estado Civil Asociado | 1 |
| Cant.CréditosVigentes | 1 |
| AP | 0 |
| **Probabilidad** | **0.000%** |
| Categoría | **Categoría 4** |
| Calificación | A |

| % | Categoría |
|---|---|
| 78.40% | Categoría 1 |
| 39.20% | Categoría 2 |
| 25.80% | Categoría 3 |
| 0.00% | Categoría 4 |
| 95.20% | Categoría 5 |

**Rangos Calificación estimación Modelo Logit**

| Calificación | Desde | Hasta | Calificación |
|---|---|---|---|
| A | 0.000% | 9.328% | A |
| B | 9.328% | 16.218% | B |
| C | 16.218% | 23.109% | C |
| D | 23.109% | 30.000% | D |
| E | 30.000% | 100.000% | E |

Fuente: Elaboración propia, 2022.